

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

26.11.2004

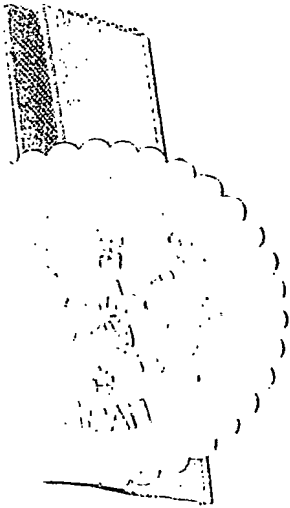
別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日            2 0 0 3 年 1 1 月 1 2 日  
Date of Application:

出 願 番 号            特 願 2 0 0 3 - 3 8 3 0 7 2  
Application Number:  
[ST. 10/C] :            [ J P 2 0 0 3 - 3 8 3 0 7 2 ]

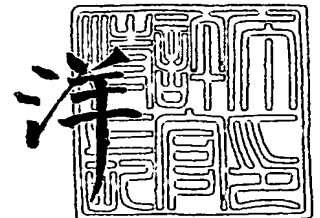
出      願      人            本 田 技 研 工 業 株 式 有 限 公 司  
Applicant(s):



2 0 0 5 年    1 月    6 日

特許庁長官  
Commissioner,  
Japan Patent Office

小 川



BEST AVAILABLE COPY

【書類名】 特許願  
【整理番号】 H103234301  
【提出日】 平成15年11月12日  
【あて先】 特許庁長官 殿  
【国際特許分類】 G10L 15/00  
【発明者】  
    【住所又は居所】 埼玉県和光市中央1丁目4番1号  
                        株式会社本田技術研究所内  
    【氏名】 中臺 一博  
【発明者】  
    【住所又は居所】 埼玉県和光市中央1丁目4番1号  
                        株式会社本田技術研究所内  
    【氏名】 奥乃 博  
【発明者】  
    【住所又は居所】 埼玉県和光市中央1丁目4番1号  
                        株式会社本田技術研究所内  
    【氏名】 辻野 広司  
【特許出願人】  
    【識別番号】 000005326  
    【氏名又は名称】 本田技研工業株式会社  
【代理人】  
    【識別番号】 100064414  
    【弁理士】  
    【氏名又は名称】 磯野 道造  
    【電話番号】 03-5211-2488  
【手数料の表示】  
    【予納台帳番号】 015392  
    【納付金額】 21,000円  
【提出物件の目録】  
    【物件名】 特許請求の範囲 1  
    【物件名】 明細書 1  
    【物件名】 図面 1  
    【物件名】 要約書 1  
    【包括委任状番号】 9713945

**【書類名】 特許請求の範囲****【請求項 1】**

複数のマイクが検出した音響信号から、特定の話者の音声認識して文字情報に変換する音声認識装置であって、

前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、

前記音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、

音源分離部が分離した音声信号に基づき、その音声信号の特徴を抽出する特徴抽出部と

、断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、

前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、

前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部と、を備えることを特徴とする音声認識装置。

**【請求項 2】**

前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成されたことを特徴とする請求項 1 に記載の音声認識装置。

**【請求項 3】**

前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されたことを特徴とする請求項 1 又は請求項 2 に記載の音声認識装置。

**【請求項 4】**

前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、

前記線形和に使用する重みが、学習により決定されたことを特徴とする請求項 1 から請求項 3 のいずれか 1 項に記載の音声認識装置。

**【請求項 5】**

前記話者を特定する話者同定部をさらに備え、

前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、

前記音響モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されたことを特徴とする請求項 1 から請求項 4 のいずれか 1 項に記載の音声認識装置。

【書類名】明細書

【発明の名称】音声認識装置

【技術分野】

【0001】

本発明は、音声認識装置に関し、詳しくは、話者や、音声認識装置を備えた移動体が移動しても高い精度で認識可能な音声認識装置に関する。

【背景技術】

【0002】

近年、音声認識技術は、実用化の域に入ってきており、情報の音声入力などに利用され始めている。一方、ロボットの研究開発も盛んとなっており、音声認識技術は、ロボットを実用化するための一つのキー技術ともなっている。すなわち、ロボットと人間との知的なソーシャルインタラクションを行うためには、人間の言葉をロボットが理解する必要があるため、音声認識の精度が重要となっている。

【0003】

ところが、実際に人とのコミュニケーションを行うためには、実験室において口元に設置したマイクで音声を入力して行う音声認識とは異なるいくつかの問題がある。

例えば、実際の環境には様々な雑音があり、雑音の中から必要な音声信号を抽出しなければ音声認識をすることができない。また、話者が複数存在する場合にも、同様に認識の対象とする話者の音声のみを抽出する必要がある。また、音声認識においては、一般に隠れマルコフモデル (HMM: Hidden Markov Model) というモデルを利用して内容を特定するが、話者の位置 (音響認識装置のマイクを基準とした方向) が異なると、話者の声の聞こえ方も異なることから、認識率に影響を及ぼすという問題がある。

【0004】

このようなことから、本発明者を含む研究グループでは、アクティブオーディションにより複数の音源の定位・分離・認識を行う技術を発表している (非特許文献1参照)。

この技術は、人間の耳に相当する位置に2つのマイクを配置し、複数の話者が同時に発話した場合に、一人の発した単語を認識する技術である。詳しくは、2本のマイクから入力された音響信号から、話者の位置を定位し、各話者の音声を分離した上で、音声認識する。この認識の際、移動体 (音声認識装置を備えたロボット等) から見て $-90^{\circ}$  から  $90^{\circ}$  まで  $10^{\circ}$  おきの方向に対する各話者の音響モデルを予め作成しておく。そして、認識時には、それらの音響モデルを用いて並列に認識プロセスを実行する。

【非特許文献1】 "A Humanoid Listens to three simultaneous talkers by Integrating Active Audition and Face Recognition" Kazuhiro Nakadai, et al., IJCAI-03 Workshop on Issues in Designing Physical Agents for Synaptic Real-Time Environments: World Modeling, Planning, Learning and Communicating, PP117-124

【発明の開示】

【発明が解決しようとする課題】

【0005】

しかしながら、前記した従来技術では、話者や移動体が移動する場合には、その都度移動体に対する話者の位置が変化するため、予め用意された音響モデルの方向と異なる方向に話者が位置すると、認識率が低下するという問題があった。

本発明は、このような背景に鑑みてなされたもので、話者や、移動体が移動しても高い精度で認識可能な音声認識装置を提供することを課題とする。

【課題を解決するための手段】

【0006】

前記課題を解決するため、本発明の音声認識装置は、複数のマイクが検出した音響信号から、特定の話者の音声を認識して文字情報に変換する音声認識装置であって、前記複数のマイクが検出した音響信号に基づき、前記特定の話者の音源方向を特定する音源定位部と、前記音源定位部が特定した音源方向に基づき、前記特定の話者の音声信号を前記音響信号から分離する音源分離部と、音源分離部が分離した音声信号に基づき、その音声信号

の特徴を抽出する特徴抽出部と、断続的な複数の方向に対応した方向依存音響モデルを記憶した音響モデル記憶部と、前記音源定位部が特定した音源方向の音響モデルを、前記音響モデル記憶部の方向依存音響モデルに基づいて合成し、前記音響モデル記憶部へ記憶させる音響モデル合成部と、前記音響モデル合成部が合成した音響モデルを使用して、前記特徴抽出部が抽出した特徴について音声認識を行い、文字情報に変換する音声認識部とを備えるように構成した。

#### 【0007】

このような音声認識装置によれば、音源定位部が音源方向を特定し、音源分離部は、音源定位部が特定した音源方向の音声のみを分離する。そして、音響モデル合成部は、音源方向と、方向依存音響モデルとに基づき、その方向に適した音響モデルを合成し、音声認識部がこの音響モデルを使用して音声認識を行う。

なお、音源分離部が出力する音声信号というのは、音声としての意味を持つ情報であればよく、音声のアナログ信号そのものに限らず、デジタル化、符号化した信号や、周波数分析したスペクトルのデータを含む。

#### 【0008】

また、前記した音声認識装置では、前記音源定位部は、前記マイクが検出した音響信号を周波数分析した後、調波構造を抽出し、複数のマイクから抽出された調波構造の音圧差と位相差とを求め、この音圧差と位相差のそれぞれから音源方向の確からしさを求め、最も確からしい方向を音源方向と判断するよう構成することができる。

#### 【0009】

さらに、前記した音声認識装置では、前記音源分離部は、前記音源定位部が特定した音源方向が、前記複数のマイクの配置により決定される正面に近い場合には、狭い方向帯域の音声を分離し、正面から離れると広い方向帯域の音声を分離するアクティブ方向通過型フィルタを用いて音声分離を行うよう構成されるのが好ましい。

#### 【0010】

また、前記した音声認識装置では、前記音響モデル合成部は、前記音響モデル記憶部の方向依存音響モデルの重み付き線形和により前記音源方向の音響モデルを合成するよう構成され、前記線形和に使用する重みが、学習により決定されるのが好ましい。

#### 【0011】

また、前記した音声認識装置では、前記話者を特定する話者同定部をさらに備え、前記音響モデル記憶部は、前記話者ごとに方向依存音響モデルを有し、前記音響モデル合成部は、前記話者同定部が特定した話者の方向依存音響モデルと、音源定位部が特定した音源方向とに基づき、前記音源方向の音響モデルを前記音響モデル記憶部の方向依存音響モデルに基づいて求め、前記音響モデル記憶部へ記憶させるよう構成されるのが好ましい。

#### 【0012】

また、前記特徴抽出部で抽出された特徴、または前記音源分離部が分離した音声信号について、予め用意した雛形と比較し、前記雛形との違いが予め設定した閾値より大きい領域、例えば周波数領域や、サブバンドを同定し、同定された領域については、その特徴としての信頼性が低いことを示す指標を前記音声認識部へ出力するマスキング部をさらに備えるのが望ましい。

#### 【発明の効果】

#### 【0013】

本発明の音声認識装置によれば、任意の方向から発された音響信号の音源方向を特定し、その音源方向に適した音響モデルを使用して音声認識をするので、音声認識率を向上することができる。

#### 【発明を実施するための最良の形態】

#### 【0014】

#### 【第1実施形態】

次に、本発明の実施形態について、適宜図面を参照しながら詳細に説明する。図1は、本発明の実施形態に係る音声認識装置のブロック構成図である。

図1に示すように、実施形態に係る音声認識装置1は、2つのマイク $M_R$ ,  $M_L$ と、マイク $M_R$ ,  $M_L$ が検出した音響信号から、話者(音源)の位置を特定する音源定位部10と、音源定位部10が特定した音源方向及び音源定位部10で求めたスペクトルに基づいて、特定の方向の音源から来る音響を分離する音源分離部20と、複数の方向についての音響モデルを記憶した音響モデル記憶部49と、音響モデル記憶部49内の音響モデル及び音源定位部10が特定した音源方向に基づいて、その音源方向の音響モデルを合成する音響モデル合成部40と、音源分離部20が分離した特定音源のスペクトルから音響の特徴を抽出する特徴抽出部30と、音響モデル合成部40が合成した音響モデルと、特徴抽出部30が抽出した音響の特徴に基づき音声認識を行う音声認識部50とを備える。

本発明では、音響モデル合成部40が生成した、音源の方向に適した音響モデルを利用して音声認識部50が音声認識を行うため、高い認識率が実現される。

#### 【0015】

次に、実施形態に係る音声認識装置の構成要素であるマイク $M_R$ ,  $M_L$ 、音源定位部10、音源分離部20、特徴抽出部30、音響モデル合成部40、及び音声認識部50についてそれぞれ説明する。

#### 【0016】

##### 《マイク $M_R$ , $M_L$ 》

マイク $M_R$ ,  $M_L$ は、音を検出して電気信号として出力する一般的なマイクである。本実施形態では、2つとしているが、複数であれば幾つでもよく、例えば3つ、4つを使用しても構わない。マイク $M_R$ ,  $M_L$ は、例えば、移動体であるロボットRBの両耳の部分に設けられる。

マイク $M_R$ ,  $M_L$ の配置は、音声認識装置1の正面を決定する。すなわち、マイク $M_R$ ,  $M_L$ の集音方向のベクトルの和の方向が音声認識装置1の正面となる。図1に示すように、ロボットRBの頭の左右両脇にマイク $M_R$ ,  $M_L$ が1つずつ設けられていれば、ロボットRBの正面が音声認識装置1の正面となる。

#### 【0017】

##### 《音源定位部10》

図2は、音源定位部の一例を示すブロック構成図であり、図3及び図4は、音源定位部の動作を説明する図である。

音源定位部10は、2つのマイク $M_R$ ,  $M_L$ から入力された2つの音響信号から、各話者 $HM_n$ (図3では、 $HM_1$ ,  $HM_2$ )の音源方向を定位する。音源定位方法は、マイク $M_R$ ,  $M_L$ に入力された音響信号の位相差を利用する方法、ロボットRBの頭部伝達関数を用いて推定する方法、右と左のマイク $M_R$ ,  $M_L$ から入力された信号の相互相関をとる方法などがあり、それぞれ精度を上げるため、種々の改良が加えられているが、ここでは、本発明者が改良した手法を例にして説明する。

#### 【0018】

音源定位部10は、図2に示すように、周波数分析部11、ピーク抽出部12、調波構造抽出部13、IPD計算部14、IID計算部15、聴覚エビポラ幾何仮説データ16、確信度計算部17、及び確信度統合部18を備える。

これらの各部を、図3及び図4を参照しながら説明する。場面として、ロボットRBに対し、2人の話者 $HM_1$ ,  $HM_2$ が同時に話しかける場合で説明する。

#### 【0019】

##### 〈周波数分析部11〉

周波数分析部11は、ロボットRBが備える左右のマイク $M_R$ ,  $M_L$ が検出した左右の音響信号 $CR_1$ ,  $CL_1$ から、微小時間 $\Delta t$ の時間長の信号区間を切り出し、左右のチャンネルごとにFFT(高速フーリエ変換)により周波数分析を行う。

例えば、右のマイク $M_R$ からの音響信号 $CR_1$ より得られる分析結果がスペクトル $CR_2$ であり、左のマイク $M_L$ からの音響信号 $CL_1$ より得られる分析結果がスペクトル $CL_2$ である。

なお、周波数分析は、バンドパスフィルタなど、他の手法を用いることもできる。

## 【0020】

## ＜ピーク抽出部12＞

ピーク抽出部12は、スペクトルCR2, CL2から左右のチャンネルごとに一連のピークを抽出する。ピークの抽出は、スペクトルのローカルピークをそのまま抽出するか、スペクトラルサブトラクション法に基づいた方法 (S.F.Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech, Proceedings of 1979 International conference on Acoustics, Speech, and signal Processing (ICASSP-79) 参照) で行う。後者の方法は、スペクトルからピークを抽出し、これをスペクトルから減算し、残差スペクトルを生成する。そして、その残差スペクトルからピークが見つからなくなるまでピーク抽出の処理を繰り返す。

前記スペクトルCR2, CL2に対し、ピークの抽出を行うと、例えばピークスペクトルCR3, CL3のようにピークを構成するサブバンドの信号のみが抽出される。

## 【0021】

## ＜調波構造抽出部13＞

調波構造抽出部13は、音源が有する調波構造に基づき、左右のチャンネルごとに特定の調波構造を有するピークをグループにする。例えば、人の声であれば、特定の人の声は、基本周波数の音と、基本周波数の倍音とからなるが、人により基本周波数が微妙に異なるので、その周波数の差により、複数の人の声をグループ分けすることができる。調波構造に基づいて同じグループに分けられたピークは、同じ音源から発せられた信号と推定できる。例えば、複数(j人)の話者が同時に話していれば、複数(j個)の調波構造が抽出される。

## 【0022】

図3においては、ピークスペクトルCR3, CL3の、ピークP1, P3, P5を一つのグループにして調波構造CR41, CL41とし、ピークP2, P4, P6を一つのグループにして調波構造CR42, CL42としている。

## 【0023】

## ＜IPD計算部14＞

IPD計算部14は、調波構造抽出部13が抽出した調波構造CR41, CR42, CL41, CL42のスペクトルから、IPD(両耳間位相差)を計算する部分である。

IPD計算部14は、調波構造j(例えば、調波構造CR41)に含まれているピーク周波数 $f_k$ の倍音に対応するスペクトルのサブバンドを、右と左の両チャンネル(例えば、調波構造CR41と調波構造CL41)から選択し、次式(1)により $IPD \Delta \phi(f_k)$ を計算する。調波構造CR41と調波構造CL41から計算した $IPD \Delta \phi(f_k)$ は、例えば図4に示す両耳間位相差C51のようになる。

## 【0024】

## 【数1】

$$\Delta \phi(f_k) = \arctan \left( \frac{\Im[S_r(f_k)]}{\Re[S_r(f_k)]} \right) - \arctan \left( \frac{\Im[S_l(f_k)]}{\Re[S_l(f_k)]} \right) \dots (1)$$

但し、

$\Delta \phi(f_k)$  :  $f_k$ のIPD(両耳間位相差)

$\Im[S_r(f_k)]$  : 右の入力信号のピーク $f_k$ のスペクトル虚部

$\Re[S_r(f_k)]$  : 右の入力信号のピーク $f_k$ のスペクトル実部

$\Im[S_l(f_k)]$  : 左の入力信号のピーク $f_k$ のスペクトル虚部

$\Re[S_l(f_k)]$  : 左の入力信号のピーク $f_k$ のスペクトル実部

## 【0025】

## ＜IID計算部15＞

IID計算部15は、各調波構造にある各倍音について、左のマイクM<sub>L</sub>から入力され

た音の音圧と、右のマイク  $M_R$  から入力された音の音圧との差（両耳間音圧差）を計算する部分である。

I I D 計算部 15 は、調波構造  $j$ （例えば、調波構造 C R 4 1, C L 4 1）に含まれているピーク周波数  $f_k$  の倍音に対応するスペクトルのサブバンドを、右と左の両チャンネル（例えば、調波構造 C R 4 1 と調波構造 C L 4 1）から選択し、次式（2）により I I D  $\Delta \rho(f_k)$  を計算する。調波構造 C R 4 1 と調波構造 C L 4 1 から計算した I I D  $\Delta \rho(f_k)$  は、例えば図 4 に示す両耳間音圧差 C 6 1 のようになる。

【0026】

【数2】

$$\Delta \rho(f_k) = p_r(f_k) - p_l(f_k) \quad \dots (2)$$

但し、

$\Delta \rho(f_k)$  :  $f_k$  の I I D（両耳間音圧差）

$p_r(f_k)$  : 右の入力信号のピーク  $f_k$  のパワー

$p_l(f_k)$  : 左の入力信号のピーク  $f_k$  のパワー

$$p_r(f_k) = 10 \log_{10} (\Im[S_r(f_k)]^2 + \Re[S_r(f_k)]^2)$$

$$p_l(f_k) = 10 \log_{10} (\Im[S_l(f_k)]^2 + \Re[S_l(f_k)]^2)$$

【0027】

<聴覚エビポーラ幾何仮説データ 16>

聴覚エビポーラ幾何仮説データ 16 は、図 5 に示すように、ロボット R B の頭部を想定した球体を上から見たときに、音源 S と、ロボット R B の両耳のマイク  $M_R$ ,  $M_L$  との距離差から生じる時間差に基づき想定される位相差のデータである。

聴覚エビポーラ幾何により、位相差  $\Delta \phi$  は、次式（3）により求められる。ここでは、頭部形状を球と仮定している。

【0028】

【数3】

$$\Delta \phi = \frac{2\pi f}{v} \times r(\theta + \sin \theta) \quad \dots (3)$$

【0029】

ここで、 $\Delta \phi$  は両耳間位相差（I P D）、 $v$  は音速、 $f$  は周波数、 $r$  は両耳間の距離  $2r$  から求まる値、 $\theta$  は音源方向を示す。

式（3）により、各音源方向より発せられた音響信号の周波数  $f$  と位相差  $\Delta \phi$  の関係は、図 6 のようになる。

【0030】

<確信度計算部 17>

確信度計算部 17 は、I P D 及び I I D のそれぞれの確信度を計算する。

— I P D 確信度—

I P D の確信度は、調波構造  $j$ （例えば、調波構造 C R 4 1, C L 4 1）が含まれている倍音  $f_k$  がどの方向から来ているらしいかを  $\theta$  の関数として求め、これを確率関数にあてはめる。

まず、 $f_k$  の I P D の仮説（予想値）を次式（4）に基づき計算する。

【0031】

【数4】

$$\Delta \phi_h(\theta, f_k) = \frac{2\pi f_k}{v} \times r(\theta + \sin \theta) \quad \dots (4)$$

【0032】



$\Delta \phi_h(\theta, f_k)$ は、ある調波構造内の $k$ 番目の倍音 $f_k$ に対して音源方向が $\theta$ の場合のIPDの仮説(予想値)を示す。IPDの仮説は、例えば音源方向 $\theta$ を、 $\pm 90^\circ$ の範囲で $5^\circ$ おきに変化させて計37個の仮説を計算する。もっとも、より細かい角度ごとに計算しても、より大まかな角度ごとに計算してもかまわない。

次に、次式(5)により、 $\Delta \phi_h(\theta, f_k)$ と $\Delta \phi(f_k)$ の差を求め、すべてのピーク $f_k$ について合計する。この差は、仮説と入力との距離を表し、 $\theta$ が話者のいる方向に近いと小さく、遠いと大きくなる。

【0033】

【数5】

$$d(\theta) = \frac{1}{K} \sum_{k=0}^{K-1} \frac{(\Delta \phi_h(\theta, f_k) - \Delta \phi(f_k))^2}{f_k} \quad \dots (5)$$

【0034】

ここで、 $\Delta \phi(f_k)$ は、ある調波構造に含まれるある倍音 $f_k$ のIPDであり、 $K$ は、その調波構造に含まれる倍音の数を示す。

得られた $d(\theta)$ を、次式(6)の確率密度関数に代入し、確信度 $B_{IPD}(\theta)$ を得る。

【0035】

【数6】

$$B_{IPD}(\theta) = \int_{-\infty}^{x(\theta)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \quad \dots (6)$$

ここで、 $X(\theta) = (d(\theta) - m) / (\sqrt{s/n})$ 、 $m$ は、 $d(\theta)$ の平均、 $s$ は $d(\theta)$ の分散であり、 $n$ はIPDの仮説の個数(本実施形態では37個)である。

【0036】

—IID確信度—

IIDの確信度は、調波構造 $j$ が含む倍音の音圧差の合計を次式(7)で計算して求める。

【0037】

【数7】

$$S = \sum_{k=0}^{K-1} \Delta \rho(f_k) \quad \dots (7)$$

【0038】

ここで、 $K$ は、その調波構造に含まれる倍音の数を示し、 $\Delta \rho(f_k)$ は、IID計算部15で求めたIIDである。

次に、表1を利用して、音源方向の右らしさ、正面らしさ、左らしさを確信度とする。なお、表1は、実験的に得られた値である。

例えば、表1を参照して、仮説の音源方位 $\theta$ が $40^\circ$ で、音圧差 $S$ が正であれば確信度 $B_{IID}(\theta)$ は、左上の欄を参照して0.35とする。

【0039】

【表1】

| $\theta$ |   | $90^\circ \sim 30^\circ$ | $30^\circ \sim -30^\circ$ | $-30^\circ \sim -90^\circ$ |
|----------|---|--------------------------|---------------------------|----------------------------|
| S        | + | 0.35                     | 0.5                       | 0.65                       |
|          | - | 0.65                     | 0.5                       | 0.35                       |

【0040】

〈確信度統合部18〉

確信度統合部18は、Dempster-Shafer理論に基づき、IPDとIIDの確信度 $B_{IPD}(\theta)$ 、 $B_{IID}(\theta)$ を次式(8)によって統合し、統合確信度 $B_{IPD+IID}(\theta)$ を計算する。そして、統合確信度 $B_{IPD+IID}(\theta)$ が最も大きくなる音源方向 $\theta$ を、話者方向とする。

【0041】

【数8】

$$B_{IPD+IID}(\theta) = 1 - (1 - B_{IPD}(\theta))(1 - B_{IID}(\theta)) \quad \dots (8)$$

【0042】

以上のような聴覚エビポーラ幾何を使用した仮説に代えて、頭部伝達関数を用いた仮説データ、又は散乱理論に基づく仮説データを用いることもできる。

(頭部伝達関数仮説データ)

頭部伝達関数仮説データは、ロボット周囲から発せられたインパルスより得られる、マイク $M_R$ とマイク $M_L$ で検出した音の位相差及び音圧差である。

頭部伝達関数仮説データは、 $-90^\circ$  から  $90^\circ$  の間の適当な間隔 (例えば  $5^\circ$ ) の方向から発したインパルスを、マイク $M_R$ ,  $M_L$ で検出し、それぞれを周波数分析して周波数 $f$ に対する位相応答及び振幅応答を求め、その差を計算することによって得られる。

得られた頭部伝達関数仮説データは、図7(a)のIPD及び(b)のIIDのようになる。

頭部伝達関数を用いる場合には、IPDだけではなく、IIDについてもある音源方向から来た音の周波数とIIDの関係が求められるので、IPDとIIDの両方について距離データ $d(\theta)$ を作ってから確信度を求める。仮説データの作成方法は、IPDとIIDで変わりはない。

聴覚エビポーラ幾何を利用した仮説データの作成方法と異なり、計算ではなく計測で、各音源方向で発せられた信号に対する周波数 $f$ とIPDの関係を求める。すなわち、図7(a), (b)にある実測値から、それぞれの仮説と入力との距離である $d(\theta)$ を直接計算する。

【0043】

(散乱理論に基づく仮説データ)

散乱理論は、ロボット頭部による散乱波を考慮して、IPD、IIDの双方を計算的に推定する理論である。ここでは、ロボットの頭部を半径 $a$ の球と仮定する。また頭部の中心の座標を極座標の原点とする。

点音源の位置を $r_0$ 、観測点を $r$ とすると、観測点における直接音によるポテンシャルは、次式(9)によって定義される。

【数9】

$$V^i = \frac{v}{2\pi R f} e^{i \frac{2\pi R f}{v}} \quad \dots (9)$$

但し、

$f$ : 点音源の周波数

$v$ : 音速

$R$ : 点音源と観測点の距離

また、観測点 $r$ を頭部表面とすると、直接音と散乱音によるポテンシャルは、次式(10)で定義される。

【数10】

$$S(\theta, f) = V^i + V^s$$

$$= - \left( \frac{v}{2\pi a f} \right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos \theta) \frac{h_n^{(1)} \left( \frac{2\pi r_0}{v} f \right)}{h_n^{(1)'} \left( \frac{2\pi a}{v} f \right)} \quad \dots (10)$$

但し、

$V^S$ : 散乱音によるポテンシャル

$P_n$ : 第一種Legendre関数

$h_n^{(1)}$ : 第一種球ハンケル関数

$M_R$ の極座標を  $(a, \pi/2, 0)$ 、 $M_L$ の極座標を  $(a, -\pi/2, 0)$  とすると、それぞれにおけるポテンシャルは、次式 (11)、(12) で表される。

【数11】

$$S_L(\theta, f) = S(f, \frac{\pi}{2} - \theta) \quad \dots (11)$$

【数12】

$$S_R(\theta, f) = S(f, -\frac{\pi}{2} - \theta) \quad \dots (12)$$

従って、散乱理論に基づく位相差  $IPD \Delta \phi_s(\theta, f)$  と音圧差  $IID \Delta \rho_s(\theta, f)$  は、それぞれ次式 (14)、(15) により求められる。

【数13】

$$\Delta \phi_s(\theta, f) = \arg(S_L(\theta, f)) - \arg(S_R(\theta, f)) \quad \dots (13)$$

【数14】

$$\Delta \rho_s(\theta, f) = 20 \log_{10} \frac{|S_L(\theta, f)|}{|S_R(\theta, f)|} \quad \dots (14)$$

【0044】

そして、前記 (4) 式の  $\Delta \phi_h(\theta, f_k)$  を前記 (13) 式の  $IPD \Delta \phi_s(\theta, f)$  に置き換え、前記した聴覚エピソード幾何を用いた場合と同じ手順で  $B_{IPD}(\theta)$  を求める。

すなわち、 $\Delta \phi_s(\theta, f_k)$  と  $\Delta \phi(f_k)$  の差を求め、すべてのピーク  $f_k$  について合計して  $d(\theta)$  を求め、得られた  $d(\theta)$  を、前記式 (6) の確率密度関数に代入し、確信度  $B_{IPD}(\theta)$  を得る。

【0045】

$IID$  も  $IPD$  と同じ方法で  $d(\theta)$  と  $B_{IID}(\theta)$  を計算する。具体的には、 $\Delta \phi$  を  $\Delta \rho$  とし、前記 (4) 式の  $\Delta \phi_h(\theta, f_k)$  を前記 (14) 式の  $IPD \Delta \rho_s(\theta, f_k)$  で置き換える。そして、 $\Delta \rho_s(\theta, f_k)$  と  $\Delta \rho(f_k)$  の差を求め、すべてのピーク  $f_k$  について合計して  $d(\theta)$  を求め、得られた  $d(\theta)$  を、前記式 (6) の確率密度関数に代入し、確信度  $B_{IID}(\theta)$  を得る。

【0046】

《音源分離部20》

音源分離部20は、音源定位部10により定位された各音源方向の情報、並びに音源定位部で計算したスペクトル（例えばスペクトルCR2）により、各話者HMnの音響（音声）信号を分離する部分である。音源分離方法には、ビームフォーミング、ナルフォーミング、ピーク追跡、指向性マイク、ICA（Independent Component Analysis：独立成分分析）など、従来からある手法を用いることができるが、ここでは、本発明者が開発したアクティブ方向通過型フィルタによる方法について説明する。

音源方向の情報を利用して音源を分離する場合、音源の方向がロボットRBの正面から離れるにつれ、2本のマイクを用いて推定した音源方向情報の精度を期待できなくなる。そこで、本実施形態では、正面方向の音源については通過させる方向の範囲を狭く、正面から離れた音源では広くとるように通過帯域をアクティブに制御して、音源の分離精度を向上させる。

【0047】

具体的には、音源分離部 20 は、図 8 に示すように、通過帯域関数 21 と、サブバンド選択部 22 と、音響信号再合成部 23 とを有する。

#### 【0048】

〈通過帯域関数 21〉

通過帯域関数 21 は、図 9 に示したように、音源方向と通過帯域幅の関数で、音源方向が、正面 ( $0^\circ$ ) から離れるにつれ、方向情報の精度を期待できなくなることから、音源方向が正面から離れるほど通過帯域幅が大きくなるように予め設定した関数である。

#### 【0049】

〈サブバンド選択部 22〉

サブバンド選択部 22 は、スペクトル CR2, CL2 の各周波数の値 (これを「サブバンド」という) から、特定の方向から来たと推測されるサブバンドを選択する部分である。

サブバンド選択部 22 では、図 10 に示すように、音源定位部 10 で生成した左右の入力音のスペクトル CR2, CL2 から、各スペクトルのサブバンドについて、前記式 (1)、(2) に従い、 $IPD\Delta\phi(f_i)$  及び  $IID\Delta\rho(f_i)$  を計算する (図 10 の両耳間位相差 C52, 両耳間音圧差 C62 参照)。

そして、音源定位部 10 で得られた  $\theta_j$  を抽出すべき音源方向とし、通過帯域関数 21 を参照して、 $\theta_j$  に対応する通過帯域幅  $\delta(\theta_j)$  を取得する。取得した通過帯域幅  $\delta(\theta_s)$  を用いて、通過帯域の最大値  $\theta_h$  と最小値  $\theta_l$  を次式 (15) により求める。通過帯域 B は、方向として平面図で図示すると、例えば図 11 のようになる。

#### 【0050】

【数 15】

$$\left. \begin{aligned} \theta_l &= \theta_j - \delta(\theta_j) \\ \theta_h &= \theta_j + \delta(\theta_j) \end{aligned} \right\} \dots (15)$$

#### 【0051】

次に、 $\theta_l$  と  $\theta_h$  に対応する IPD と IID を推定する。これらの推定には、予め計測、又は計算した伝達関数を利用する。伝達関数は、音源方向  $\theta$  から来る信号に対して周波数  $f$  と IPD、IID をそれぞれ関係づけている関数で、前記したエピソード幾何や、頭部伝達関数、散乱理論などを用いる。推定した IPD は、例えば図 10 の両耳間位相差 C53 における  $\Delta\phi_l(f)$ ,  $\Delta\phi_h(f)$  であり、推定した IID は、例えば図 10 の両耳間音圧差 C63 における  $\Delta\rho_l(f)$ ,  $\Delta\rho_h(f)$  である。

#### 【0052】

次に、音源方向  $\theta_s$  に対して、ロボット RB の伝達関数を利用して、入力スペクトルの周波数  $f$  に応じ、周波数  $f$  が所定の閾値周波数  $f_{th}$  より小さければ IPD によりサブバンドを選択し、大きければ IID によりサブバンドを選択する。すなわち、以下の条件式 (16) を満たすサブバンドを選択する。

#### 【0053】

【数 16】

$$\left. \begin{aligned} f < f_{th} : \Delta\phi_l(f_i) \leq \Delta\phi(f_i) \leq \Delta\phi_h(f_i) \\ f \geq f_{th} : \Delta\rho_l(f_i) \leq \Delta\rho(f_i) \leq \Delta\rho_h(f_i) \end{aligned} \right\} \dots (16)$$

#### 【0054】

ここで、 $f_{th}$  は、フィルタリングの判断基準に IPD と IID のどちらを用いるかを定める閾値周波数である。

この条件式によれば、例えば、図 10 の両耳間位相差 C53 においては、周波数  $f_{th}$  より低い周波数で、IPD が  $\Delta\phi_l(f)$  と  $\Delta\phi_h(f)$  の間にある周波数  $f_i$  のサブバンド (斜線部) が選択される。一方、図 10 の両耳間音圧差 C63 においては、周波数  $f_{th}$  より

り高い周波数で、IIDが $\Delta\rho_l(f)$ と $\Delta\rho_h(f)$ の間にあるサブバンド(斜線部)が選択される。この選択されたサブバンドからなるスペクトルを本明細書において「選択スペクトル」という。

#### 【0055】

〈音響信号再合成部23〉

音響信号再合成部23は、サブバンド選択部22が選択した右又は左の選択スペクトルから、音響信号を逆フーリエ変換することで再合成し、特定範囲にある音源から発せられた音響信号(図10の音声信号C7参照)を得る。

#### 【0056】

以上、本実施形態の音源分離部20について説明したが、音源分離の方法には、この他に指向性マイクを利用した方法がある。即ち、指向性が狭いマイクをロボットRBに設けておき、音源定位部10で得られた音源方向 $\theta_j$ の方向に指向性マイクを向けるよう、顔の向きを変えれば、その方向から来る音声だけを取得することができる。

この指向性マイクによる方法の場合、1つの指向性マイクしかない場合には、1人の音声しか取得できないという問題もあるが、複数の指向性マイクを所定角度おきに設けておき、音源方向の指向性マイクからの音声信号を利用するようにすれば、複数人の音声の同時取得も可能である。

#### 【0057】

《特徴抽出部30》

特徴抽出部30は、音源分離部20で分離された音声から音声認識に必要な特徴を抽出する部分である。音声の特徴としては、音声を周波数分析した線形スペクトルや、メル周波数ケプストラム係数(MFCC: Mel-Frequency Cepstrum Coefficient)を用いることができる。本実施形態では、MFCCを用いる場合で説明する。

#### 【0058】

特徴抽出部30は、図12に示すように、対数変換部31、メル周波数変換部32、及びコサイン変換部33を有する。

対数変換部31は、サブバンド選択部22が選択した選択スペクトルの振幅を対数に変換して、線形対数スペクトルを得る。

メル周波数変換部32は、対数変換部31が生成した線形対数スペクトルを、メル周波数のバンドパスフィルタに通し、周波数がメルスケールに変換されたメル周波数対数スペクトルを得る。

コサイン変換部33は、メル周波数変換部32が生成したメル周波数対数スペクトルをコサイン変換する。このコサイン変換により得られた係数がMFCCとなる。

#### 【0059】

また、雑音などによって入力音声が変形している場合は、そのスペクトルサブバンドを特徴として信用しないよう、指標(0から1)を付与するマスキング部を、特徴抽出部30の中または後に任意的に追加してもよい。

具体的な例としては、MFCCを求める際、まず線形スペクトルの段階で、予め用意した雛形のスペクトルと比較して、前記雛形との違いが予め設定した閾値より大きいスペクトルの領域を同定し、この領域を入力音声の変形の影響を受けているサブバンドと同定する。そして、同定したサブバンドを0倍したものと、1倍したものの2つのMFCCを求める。得られたMFCCを比較し、差が大きいものには0に近い指標 $\omega$ を、小さいものには1に近い指標 $\omega$ を付与する。または、実験的な閾値を定め、閾値以下は1、以上は0を付与する。この指標 $\omega$ は音声認識時に用いられる。

#### 【0060】

なお、指向性マイクを用いて音源分離を行う場合には、指向性マイクから得られた分離音声に対し、FFTやバンドパスフィルタなどの一般的な周波数分析手法を用いてスペクトルを得る。

#### 【0061】

《音響モデル合成部40》

音響モデル合成部 40 は、音響モデル記憶部 49 に記憶された方向依存音響モデルから、定位された各音源方位に応じた音響モデルを合成する部分である。

音響モデル合成部 40 は、図 13 に示すように、コサイン逆変換部 41 と、線形変換部 42 と、指数変換部 43 と、パラメータ合成部 44 と、対数変換部 45 と、メル周波数変換部 46 と、コサイン変換部 47 とを有し、音響モデル記憶部 49 に記憶された方向依存音響モデル  $H(\theta_n)$  を参照して  $\theta$  方向の音響モデルを合成する。

#### 【0062】

〈音響モデル記憶部 49〉

音響モデル記憶部 49 には、ロボット RB の正面を基準とした方向  $\theta_n$  ごとに、方向  $\theta_n$  に適した音響モデルである方向依存音響モデル  $H(\theta_n)$  が記憶されている。方向依存音響モデル  $H(\theta_n)$  は、特定の方向  $\theta_n$  から発せられた人物の音声の MFCC を、複数人物について隠れマルコフモデル (HMM) として学習させたものである。各方向依存音響モデル  $H(\theta_n)$  は、図 14 に示すように、例えば音素を認識単位とし、音素ごとに対応するモノフォンのサブモデル  $h(m, \theta_n)$  を記憶している。なお、サブモデルは、トライフォンで作成してもよい。

サブモデル  $h(m, \theta_n)$  の数は、例えば方向  $\theta_n$  について  $-90^\circ \sim 90^\circ$  まで  $30^\circ$  おきに 7 個のモデルを持ち、サブモデルを 40 個のモノフォンで構成しているとすれば、合計  $7 \times 40 = 280$  個となる。

サブモデル  $h(m, \theta_n)$  は、状態数、各状態の確率密度分布、状態遷移確率の各パラメータを有している。本実施形態では、各音素の状態数は、前部 (状態 1)、中間部 (状態 2)、後部 (状態 3) の 3 つに固定している。また、確率密度分布は、正規分布に固定する。したがって、本実施形態では、状態遷移確率  $P$  と、正規分布のパラメータ、つまり平均  $\mu$  及び標準偏差  $\sigma$  を学習させる。

#### 【0063】

サブモデル  $h(m, \theta_n)$  の学習データは次のようにして作成する。

ロボット RB に対し、音響モデルを作成したい方向から、特定の音素からなる音声信号を図示しないスピーカにより発する。そして、検出した音響信号を特徴抽出部 30 により MFCC に変換し、後述する音声認識部 50 で音声認識させる。すると、認識した音声、音素ごとにどのくらいの確率であるかが結果として得られるが、この結果に対し、特定の方向の特定の音素であるという教師信号を与えることで音響モデルを適応学習させる。そして、サブモデルを学習するのに十分な種類 (例えば、異なる話者) の音素や単語を学習させる。

なお、学習用音声を発する際、音響モデルを作成したい方向とは異なる方向から、別の音声をノイズとして発してもよい。この場合は、前記した音源分離部 20 により音響モデルを作成したい方向の音響のみを分離した上で、特徴抽出部 30 により MFCC に変換する。また、これらの学習は、音響モデルを不特定話者のモデルとして持たせたい場合には、不特定の話者の声で学習させればよいし、特定話者ごとにモデルを持たせたい場合には、特定話者ごとに学習させればよい。

#### 【0064】

コサイン逆変換部 41 から指数変換部 43 は、確率密度分布の MFCC を線形スペクトルに戻す。つまり、確率密度分布について、特徴抽出部 30 と逆の操作をする。

#### 【0065】

〈コサイン逆変換部 41〉

コサイン逆変換部 41 は、音響モデル記憶部 49 が記憶している方向依存音響モデル  $H(\theta_n)$  が有する MFCC についてコサイン逆変換してメル対数スペクトルを生成する。

#### 【0066】

〈線形変換部 42〉

線形変換部 42 は、コサイン逆変換部 41 により生成されたメル対数スペクトルの周波数を線形周波数に変換し、対数スペクトルを生成する。

#### 【0067】

### 〈指数変換部 4 3〉

指数変換部 4 3 は、線形変換部 4 2 により生成された対数スペクトルの強度を指数変換し、線形スペクトルを生成する。線形スペクトルは、平均  $\mu$ 、標準偏差  $\sigma$  の確率密度分布として得られる。

#### 【0068】

### 〈パラメータ合成部 4 4〉

パラメータ合成部 4 4 は、図 15 に示すように、方向依存音響モデル  $H(\theta_n)$  にそれぞれ重みをかけて、音源方向  $\theta_j$  の音響モデル  $H(\theta_j)$  を合成する。方向依存音響モデル  $H(\theta_n)$  は、それぞれ前記逆コサイン逆変換部 4 1 から指数変換部 4 3 により、線形スペクトルの確率密度分布に変換され、それぞれ、平均  $\mu_{1n}$ ,  $\mu_{2n}$ ,  $\mu_{3n}$ , 標準偏差  $\sigma_{1n}$ ,  $\sigma_{2n}$ ,  $\sigma_{3n}$ , 状態遷移確率  $P_{11n}$ ,  $P_{12n}$ ,  $P_{22n}$ ,  $P_{23n}$ ,  $P_{33n}$  のパラメータを持っている。そして、これらのパラメータを、予め学習によって求められ、音響モデル記憶部 4 9 に記憶されている重み  $w_n$  と内積して、音源方向  $\theta_j$  の音響モデルを合成する。つまり、パラメータ合成部 4 4 は、方向依存音響モデル  $H(\theta_n)$  の線形和により音源方向  $\theta_j$  の音響モデルを合成している。なお、重み  $w_n$  の学習の仕方は後述する。

#### 【0069】

$H(\theta_j)$  にあるサブモデルを合成する場合には、状態 1 の平均  $\mu_{1j}$  を次式 (17) により求める。

#### 【0070】

#### 【数 17】

$$\mu_{1j} = \frac{1}{\sum_{n=1}^N w_n} \sum_{n=1}^N w_n \mu_{1n} \quad \dots (17)$$

#### 【0071】

平均  $\mu_{2j}$ ,  $\mu_{3j}$  についても同様にして求めることができる。

#### 【0072】

また、状態 1 の標準偏差  $\sigma_{1j}$  の合成については、共分散  $\sigma_{1j}^2$  を次式 (18) により求める。

#### 【数 18】

$$\sigma_{1j}^2 = \frac{1}{\sum_{n=1}^N w_n} \sum_{n=1}^N w_n \sigma_{1n}^2 \quad \dots (18)$$

#### 【0073】

標準偏差  $\sigma_{2j}$ ,  $\sigma_{3j}$  についても同様にして求めることができる。

得られた  $\mu$  と  $\sigma$  により、確率密度分布を求めることができる。

#### 【0074】

また、状態 1 の状態遷移確率  $P_{11j}$  の合成については、次式 (19) により求める。

#### 【0075】

#### 【数 19】

$$P_{11j} = \frac{1}{\sum_{n=1}^N w_n} \sum_{n=1}^N w_n P_{11n} \quad \dots (19)$$

#### 【0076】

状態遷移確率  $P_{12j}$ ,  $P_{22j}$ ,  $P_{23j}$ ,  $P_{33j}$  についても同様にして求めることができる。

## 【0077】

次に、対数変換部45からコサイン変換部47により、確率密度分布を線形スペクトルからMFCCに変換し直す。すなわち、対数変換部45は、対数変換部31と、メル周波数変換部46は、メル周波数変換部32と、コサイン変換部47は、コサイン変換部33と同様であるので、詳細な説明を省略する。

## 【0078】

なお、正規分布によらず、混合分布による場合には、前記した平均 $\mu$ 、標準偏差 $\sigma$ の計算に代えて次式(20)により確率密度分布 $f_{1n}(x)$ を求める。

## 【0079】

## 【数20】

$$f_{1j}(x) = \frac{1}{\sum_{n=1}^N w_n} \sum_{n=1}^N w_n f_{1n}(x) \quad \dots (20)$$

## 【0080】

確率密度分布 $f_{2j}(x)$ 、 $f_{3j}(x)$ についても同様にして求めることができる。

## 【0081】

パラメータ合成部44は、このようにして得られた音響モデルを、音響モデル記憶部49に記憶させる。

なお、このような音響モデルの合成は、音声認識装置1が作動している間、パラメータ合成部44がリアルタイムに行う。

## 【0082】

〈重み $w_{nm}$ の学習〉

次に、重み $w_{nm}$  ( $n$ は方向、 $m$ は音素を示す)の学習方法について説明する。

音源方向 $\theta_j$ についての、音素/a/の重み $w_{j/a/}$ を学習する場合には、最初に適当な初期値の重みの値(ベクトル)の $w_{j/a/}$ を設定しておき、この $w_{j/a/}$ を用いて合成した音響モデル $H(\theta_j)$ で、/a/を含む適当な音素列、例えば学習データ[/a//b//c/]を認識させる試行を行う。具体的には、方向 $\theta_j$ に設置したスピーカから、[/a//b//c/]の音素を発し、これを認識させる。ここで、学習データは、一つの音素/a/自体であってもよいのであるが、音素が複数つながった音素列で学習させた方がよい学習結果が出るため、音素列を使用している。

このときの認識結果が、例えば図17である。図17では、初期値の $w_{j/a/}$ を用いた音響モデル $H(\theta_j)$ での認識結果が一行目であり、2行目以下の $H(\theta_n)$ が、方向 $\theta_n$ の方向依存音響モデル $H(\theta_n)$ を使用したときの認識結果である。例えば、音響モデル $H(\theta_j)$ での認識結果は、[/x//y//z/]であり、音響モデル $H(\theta_{-90})$ での認識結果は、[/x//y//c/]であったことを示す。

## 【0083】

1回目の試行の後、まず1音素目を見て、図17の $\theta_j$ から $\theta_{90}$ までに一致する音素が認識された場合、その方向に対応するモデルの重み $w$ を $\Delta d$ 増加させる。 $\Delta d$ は実験的に求め、ここでは0.05とする。そして、一致する音素が認識されない場合、その方向に対応するモデルの重み $w$ を $k \Delta d / (n - k)$ 減少させる。つまり、正解を出した方向依存音響モデルの重みは大きくし、正解を出さなかった方向依存音響モデルの重みは減少させる。

例えば、図17の場合では、 $H(\theta_n)$ と $H(\theta_{90})$ が一致しているので、 $w_{n/a/}$ と $w_{90/a/}$ を $\Delta d$ 増加させ、それ以外の $w$ を $2 \Delta d / (n - 2)$ 減少させる。

## 【0084】

一方、1音素目に一致する音素を認識した方向 $\theta_n$ がない場合、他の方向に対して重み $w_{j/a/}$ の成分が大きい、優勢な方向依存音響モデル $H(\theta_n)$ があれば、その方向依存音響モデル $H(\theta_n)$ の重みを $\Delta d$ 減少させ、それ以外のモデルの重みを $k \Delta d / (n - k)$ 増



加させる。つまり、どの方向依存音響モデル  $H(\theta_n)$  も正しく認識できなかったということは、現在の重みの分配が良くない可能性があるから、現在重みが優勢な方向について重みを減少させる。

優勢であるかどうかは、 $w_{nm}$  が予め定められた閾値  $w_{th}$  (ここでは 0.8 とする) より大きいかどうかで判断する。優勢な方向依存音響モデル  $H(\theta_n)$  が無ければ、最大の  $w_{nm}$  のみを  $\Delta d$  減少させ、その他の方向依存音響モデル  $H(\theta_n)$  の  $w_{nm}$  を  $\Delta d / (n-1)$  増加させる。

そして、更新された重み  $w_{j/a/}$  を用いて、前記した試行を繰り返す。

【0085】

そして、音響モデル  $H(\theta_j)$  の認識結果が、正解 (/a/) となったときに、繰り返しを終了し次の音素 /b/ の文字の認識及び学習へ移る。

なお、所定の回数 (例えば 0.5 /  $\Delta d$  回) 繰り返しても、音響モデル  $H(\theta_j)$  の認識結果が正解に至らない場合、例えば /a/ の認識がうまくいかなかった場合には、次の音素 /b/ の学習へ移り、最終的にうまく認識できた音素 (例えば音素 /b/) の  $w_{j/b/}$  の分布と同じ値で  $w_{j/a/}$  を更新する。

【0086】

このようにして学習して得られた重み  $w$  は、音響モデル記憶部 49 に記憶させる。

【0087】

《音声認識部 50》

音声認識部 50 は、音源方向  $\theta_j$  に対応して合成された音響モデル  $H(\theta_j)$  を用いて、分離された各話者  $HM_n$  の音声を認識して文字情報とし、単語辞書を参照して言葉を認識し、認識結果を出力する。この音声認識の方法は一般的な隠れマルコフモデルを利用した認識方法なので、詳細な説明は省略する。

なお、マスキング部を特徴抽出部 30 の中または後に設けて、MFCC の各サブバンドの信用度を示す指標  $\omega$  が付与されている場合には、音声認識部 50 は、次のような処理を行う。

一般的な隠れマルコフモデルを利用した認識方法では、特徴抽出部 30 において抽出された特徴  $x$  ( $x$  は特徴ベクトルを表し、本実施例では MFCC を用いている) に対するある音響モデル中の状態  $S$  の出力確率は、 $f(x|S)$  で表される。特徴抽出部 30 から指標  $\omega$  が与えられる場合は、 $x$  のうちの信用できる成分  $x_r$  を計算し、次式 (21) により出力確率を求める。

【数 21】

$$f(x_r|S) = \sum_{l=1}^L P_l f(x_r|l, S) \quad \dots (21)$$

$P_l$ : 混合分布係数

$L$ : ある状態に含まれる分布の数

$$x = x_r + x_u$$

$x_u$ :  $x$  のうち信用できない成分

$x_r$ :  $x$  のうち信用できる成分

$$x_r(i) = \omega(i) \times x(i)$$

$i$ : MFCC の次元

そして、得られた出力確率と状態遷移確率を用いて、一般的な隠れマルコフモデルを利用した認識方法と同様に認識を行う。

【0088】

以上のように構成された、音声認識装置 1 による動作を説明する。

図 1 に示すように、ロボット RB のマイク  $M_R$ ,  $M_L$  に、複数の話者  $HM_n$  (図 3 参照) の音声が入力される。

そして、マイク  $M_R$ ,  $M_L$  が検出した音響信号の音源方向が音源定位部 10 で定位される。音源定位は、前記したように周波数分析、ピーク抽出、調波構造の抽出、 $IPD \cdot IID$  の計算の後、聴覚エピソード幾何に基づいた仮説データを利用して確信度を計算する。そして、 $IPD$  と  $IID$  の確信度を統合して最も可能性が高い  $\theta_j$  を音源方向とする（図 2 参照）。

#### 【0089】

次に、音源分離部 20 で、音源方向  $\theta_j$  の音を分離する。音源分離は、通過帯域関数を利用して、音源方向  $\theta_j$  の  $IPD$  及び  $IID$  のそれぞれの上限值  $\Delta \phi_h(f)$ ,  $\Delta \rho_h(f)$  及び下限値  $\Delta \phi_l(f)$ ,  $\Delta \rho_l(f)$  を求め、前記式 (16) の条件と、この上限値、下限値の条件とから、音源方向  $\theta_j$  のスペクトルと推定されるサブバンド（選択スペクトル）を選択する。その後、選択サブバンドのスペクトルを逆 FFT により変換すれば、音声信号に変換できる。

#### 【0090】

次に、特徴抽出部 30 は、音源分離部 20 が分離した選択スペクトルを、対数変換部 31、メル周波数変換部 32、コサイン変換部 33 により MFCC に変換する。

#### 【0091】

一方、音響モデル合成部 40 は、音響モデル記憶部 49 に記憶された方向依存音響モデル  $H(\theta_n)$  と、音源定位部 10 が定位した音源方向  $\theta_j$  とから、音源方向  $\theta_j$  に適切と考えられる音響モデルを合成する。

すなわち、音響モデル合成部 40 は、方向依存音響モデル  $H(\theta_n)$  を、コサイン逆変換部 41、線形変換部 42、及び指数変換部 43 により、線形スペクトルに変換する。そして、パラメータ合成部 44 は、音源方向  $\theta_j$  の重み  $w_j$  を音響モデル記憶部 49 から読み出し、これと方向依存音響モデル  $H(\theta_n)$  との内積をとって、音源方向  $\theta_j$  の音響モデル  $H(\theta_j)$  を合成する。そして、この線形スペクトルで表された音響モデル  $H(\theta_j)$  を、対数変換部 45、メル周波数変換部 46、及びコサイン変換部 47 により MFCC で表した音響モデル  $H(\theta_j)$  に変換する。

#### 【0092】

次に、音声認識部 50 は、音響モデル合成部 40 で合成された音響モデル  $H(\theta_j)$  を利用して、隠れマルコフモデルにより音声認識を行う。

#### 【0093】

このようにして、音声認識を行った結果の例が、表 2 である。

#### 【0094】

【表 2】

|          | 従来手法 |      |      |     |     |     |     | 本発明 |
|----------|------|------|------|-----|-----|-----|-----|-----|
| 音響モデルの方向 | -90° | -60° | -30° | 0   | 30° | 60° | 90° | 40° |
| 孤立単語認識率  | 20%  | 20%  | 38%  | 42% | 60% | 59% | 50% | 78% |

#### 【0095】

表 2 に示すように、方向依存音響モデルを  $-90^\circ \sim 90^\circ$  まで  $30^\circ$  おきに用意して、各音響モデルで  $40^\circ$  の方向から孤立単語を認識させたところ（従来手法）、最も認識率が高くて  $30^\circ$  方向の方向依存音響モデルを用いた  $60\%$  であった。これに対し、本実施形態の手法を使用して  $40^\circ$  方向の音響モデルを合成して、これを用いて孤立単語を認識させたところ、 $78\%$  の高い認識率を示した。このように、本実施形態の音声認識装置 1 によれば、任意の方向から音声が発せられた場合であっても、その方向に適した音響モデルをその都度合成するので、高い認識率を実現することができる。また、任意の方向の音声を認識できることから、移動している音源からの音声認識や、移動体（ロボット R B）自身が移動しているときにも、高い認識率での音声認識が可能である。

#### 【0096】

また、方向依存音響モデルを、断続的な数個、例えば音源方向にして  $60^\circ$  ごとや  $30^\circ$  ごとに記憶しておけば良く、音響モデルの学習に必要なコストを小さくすることができ

る。

さらに、合成した音響モデル一つについて音声認識を行えば良いため、複数方向の音響モデルについて音声認識を試みる並列処理も不要であり、計算コストを小さくすることができる。そのため、実時間処理や、組み込み用途には好適である。

#### 【0097】

以上、本発明の第1実施形態について説明したが、本発明は第1実施形態には限定されず、以下の実施形態のように変形して実施することが可能である。

#### 【0098】

##### [第2実施形態]

第2実施形態では、第1実施形態の音源定位部10に代えて、相互相関のピークを用いて音源方向を定位する音源定位部110を備える。なお、他の部分については第1実施形態と同様であるので説明を省略する。

#### 《音源定位部110》

第2実施形態に係る音源定位部110は、図18に示すように、フレーム切り出し部111、相互相関計算部112、ピーク抽出部113、方向推定部114を有する。

#### 【0099】

##### 〈フレーム切り出し部111〉

フレーム切り出し部111は、左右のマイク $M_R$ ,  $M_L$ に入力されたそれぞれの音響信号について、所定の時間長、例えば100msecで切り出す処理を行う。切り出し処理は、適当な時間間隔、例えば30msecごとに行われる。

#### 【0100】

##### 〈相互相関計算部112〉

相互相関計算部112は、フレーム切り出し部111が切り出した右マイク $M_R$ の音響信号と、左マイク $M_L$ の音響信号とで、次式(22)により相互相関を計算する

#### 【数22】

$$CC(T) = \int x_L(t)x_R(t+T)dt \quad \dots (22)$$

但し、

$CC(T)$  :  $x_L(t)$  と  $x_R(t)$  の相互相関

$T$  : フレーム長

$x_L(t)$  : フレーム長 $T$ で切り出された、マイク $L$ からの入力信号

$x_R(t)$  : フレーム長 $T$ で切り出された、マイク $R$ からの入力信号

#### 【0101】

##### 〈ピーク抽出部113〉

ピーク抽出部113は、得られた相互相関の結果からピークを抽出する。抽出するピークの数、音源の数が予め分かっている場合は、その数に対応したピークを大きいものから選択する。音源数が不明なときは、予め定めた閾値を超えたピークを全て抽出するか、あるいは予め定めた所定数のピークを大きいものから順に選択する。

#### 【0102】

##### 〈方向推定部114〉

音源方向 $\theta_j$ は、得られたピークから、右マイク $M_R$ と左マイク $M_L$ に入力された音響信号の到達時間差 $D$ に音速 $v$ を掛けて、図19に示す距離差 $d$ を計算し、さらに、次式により求める。

$$\theta_j = \arcsin(d/b)$$

#### 【0103】

このような相互相関を用いた音源定位部110によっても、音源方向 $\theta_j$ の方向が推定され、前記した音響モデル合成部40により、音源方向 $\theta_j$ に適した音響モデルを合成することで、認識率の向上を図ることができる。

#### 【0104】

**[第3実施形態]**

第3実施形態では、第1実施形態に加えて、音源定位部音源が同一音源から来ていることを確認しながら音声認識を行う機能を追加している。なお、第1実施形態と同じ部分については、同じ符号を付して説明を省略する。

第3実施形態に係る音声認識装置100は、第1実施形態の音声認識装置1に加え、音源定位部10が定位した音源方向を入力されて、音源を追跡し、同じ音源から音響が来続けているかを確認し、確認ができたなら、音源方向を音源分離部20へ出力するストリーム追跡部60を有している。

**【0105】**

図21に示すように、ストリーム追跡部60は、音源方向履歴記憶部61と、予測部62と、比較部63とを有する。

**【0106】**

音源方向履歴記憶部61は、図22に示すような、時間と、その時間において認識された音源の方向及び音源のピッチ（その音源の調波構造が持つ最小周波数 $f_1$ ）とが関連づけて記憶されている。

**【0107】**

予測部62は、音源方向履歴記憶部61から、直前まで追跡していた音源の音源方向の履歴を読み出し、直前までの履歴からカルマンフィルタなどにより現時点 $t_1$ での音源方向 $\theta_j$ 及びピッチ $f_1$ とからなるストリーム特徴ベクトル $(\theta_j, f_1)$ を予測し、比較部63へ出力する。

**【0108】**

比較部63は、音源定位部10から、音源定位部10で定位された現時点 $t_1$ の各話者 $j$ の音源方向 $\theta_j$ と、その音源のピッチ $f_1$ とが入力される。そして、予測部62から入力された予測したストリーム特徴ベクトル $(\theta_j, f_1)$ と、音源定位部10で定位された音源方向及びピッチから求まるストリーム特徴ベクトル $(\theta_j, f_1)$ を比較して、その差（距離）が予め定めた閾値よりも小さい場合に、音源方向 $\theta_j$ を音源分離部に出力する。また、ストリーム特徴ベクトル $(\theta_j, f_1)$ を音源方向履歴記憶部61へ記憶させる。

前記した差（距離）が、予め定めた閾値よりもより大きい場合には、定位した音源方向 $\theta_j$ を音源分離部20へ出力しないので、音声認識は行われない。なお、音源方向 $\theta_j$ とは別に、音源の追跡ができているか否かを示すデータを、比較部63から音源分離部20へ出力してもよい。

なお、ピッチ $f_1$ を用いず、音源方向 $\theta_j$ だけで予測してもよい。

**【0109】**

このようなストリーム追跡部60を有する音声認識装置100によれば、音源定位部10で音源方向が定位され、ストリーム追跡部60へ音源方向とピッチが入力される。ストリーム追跡部60では、予測部62が、音源方向履歴記憶部61に記憶された音源方向の履歴を読み出して現時点 $t_1$ でのストリーム特徴ベクトル $(\theta_j, f_1)$ を予測する。比較部63は、予測部62で予測されたストリーム特徴ベクトル $(\theta_j, f_1)$ と、音源定位部10から入力された値から求まるストリーム特徴ベクトル $(\theta_j, f_1)$ とを比較して、その差（距離）が所定の閾値より小さければ、音源方向を音源分離部20へ出力する。

音源分離部20は、音源定位部10から入力されたスペクトルのデータと、ストリーム追跡部60が出力した音源方向 $\theta_j$ のデータに基づき、第1実施形態と同様にして音源を分離する。そして、以下、特徴抽出部30、音響モデル合成部40、音声認識部50でも、第1実施形態と同様にして、処理を行う。

**【0110】**

このように、本実施形態の音声認識装置100は、音源が追跡できているか否かを確認した上で音声認識を行うので、音源が移動している場合にも、同じ音源が発し続けている音声を持続して認識するため、誤認識の可能性を低くすることができる。特に、複数の移動する音源があって、それらの音源が交差する場合などに好適である。

また、音源方向を記憶、予測していることから、その方向の所定範囲についてのみ音源

を探索すれば、処理を少なくすることができる。

【0111】

以上、本発明の実施形態について説明したが、本発明は、前記した実施形態には限定されず

例えば、音声認識装置1が、カメラと、公知の画像認識装置を有し、話者の顔を認識して、誰が話しているかを自己が有するデータベースから話者を特定する話者同定部を備え、前記方向依存音響モデルを話者ごとに有していれば、話者に適した音響モデルを合成することができるので、認識率をより高くする事ができる。あるいは、カメラを使わず、ベクトル量子化(VQ)を用いて、予め登録してある話者の音声をベクトル化したものと、音源分離部20で分離された音声をベクトル化したものとを比較し、最も距離の近い話者を結果として出力することで話者を同定してもよい。

【図面の簡単な説明】


【0112】

- 【図1】 本発明の実施形態に係る音声認識装置のブロック構成図である。
- 【図2】 音源定位部の一例を示すブロック構成図である。
- 【図3】 音源定位部の動作を説明する図である。
- 【図4】 音源定位部の動作を説明する図である。
- 【図5】 聴覚エビポーラ幾何を説明する図である。
- 【図6】 位相差 $\Delta\phi$ と周波数 $f$ の関係を示すグラフである。
- 【図7】 頭部伝達関数の一例を示すグラフである。
- 【図8】 音源分離部の一例を示すブロック構成図である。
- 【図9】 通過帯域関数の一例を示すグラフである。
- 【図10】 サブバンド選択部の動作を説明する図である。
- 【図11】 通過帯域の一例を図示した平面図である。
- 【図12】 特徴抽出部の一例を示すブロック構成図である。
- 【図13】 音響モデル合成部の一例を示すブロック構成図である。
- 【図14】 方向依存音響モデルの認識単位とサブモデルを示した図である。
- 【図15】 パラメータ合成部の動作を説明する図である。
- 【図16】 重み $w_n$ の一例を示すグラフである。
- 【図17】 重み $w$ の学習方法を説明する図である。
- 【図18】 第2実施形態に係る音声認識装置のブロック図である。
- 【図19】 音響の入力距離差を示す図である。
- 【図20】 第3実施形態に係る音声認識装置のブロック図である。
- 【図21】 ストリーム追跡部のブロック図である。
- 【図22】 音源方向の履歴を図示したグラフである。

【符号の説明】

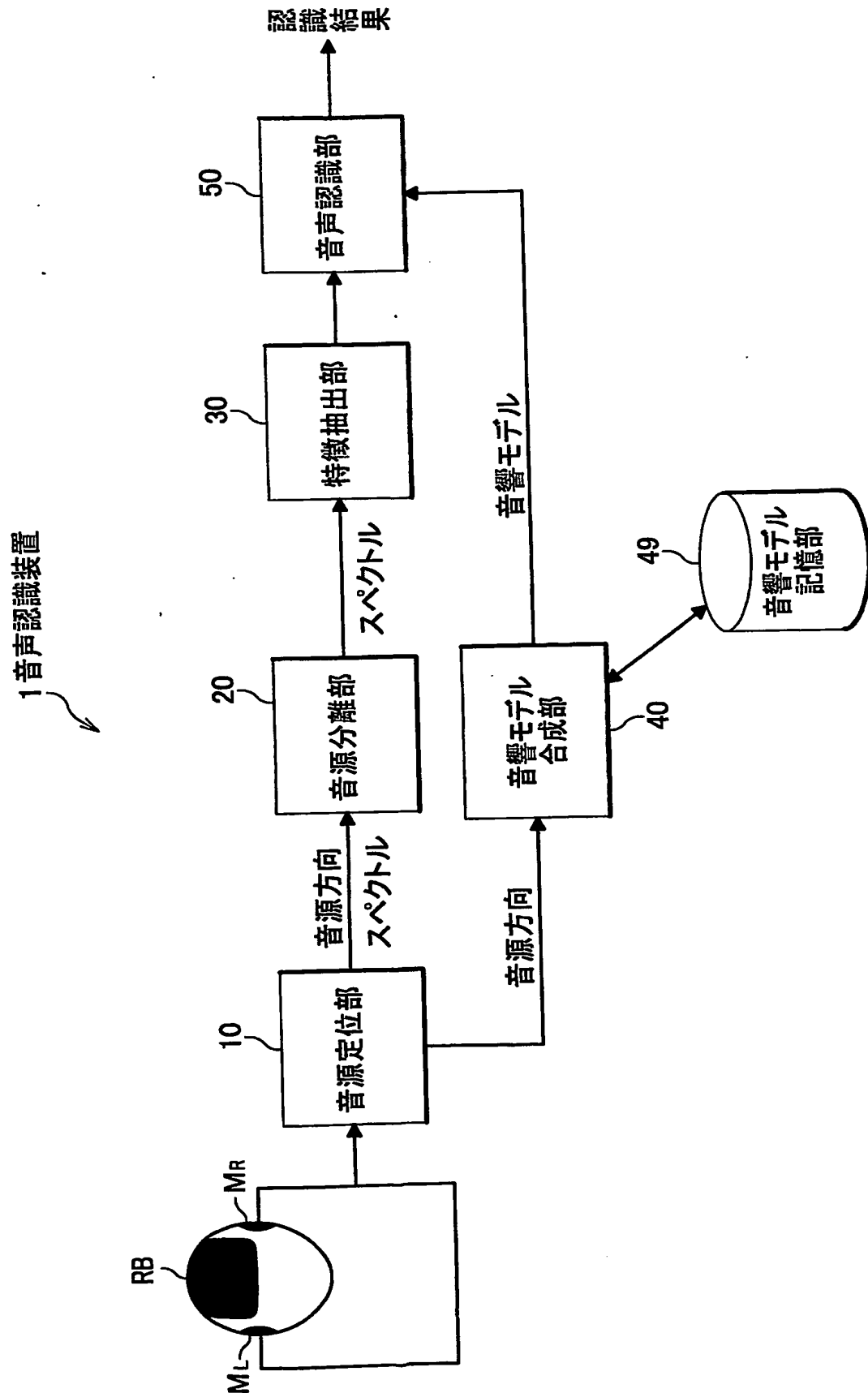
【0113】

- 1 音声認識装置
- 10 音源定位部
- 11 周波数分析部
- 12 ピーク抽出部
- 13 調波構造抽出部
- 14 IPD計算部
- 15 IID計算部
- 16 聴覚エビポーラ幾何仮説データ
- 17 確信度計算部
- 18 確信度統合部
- 20 音源分離部
- 21 通過帯域関数
- 22 サブバンド選択部

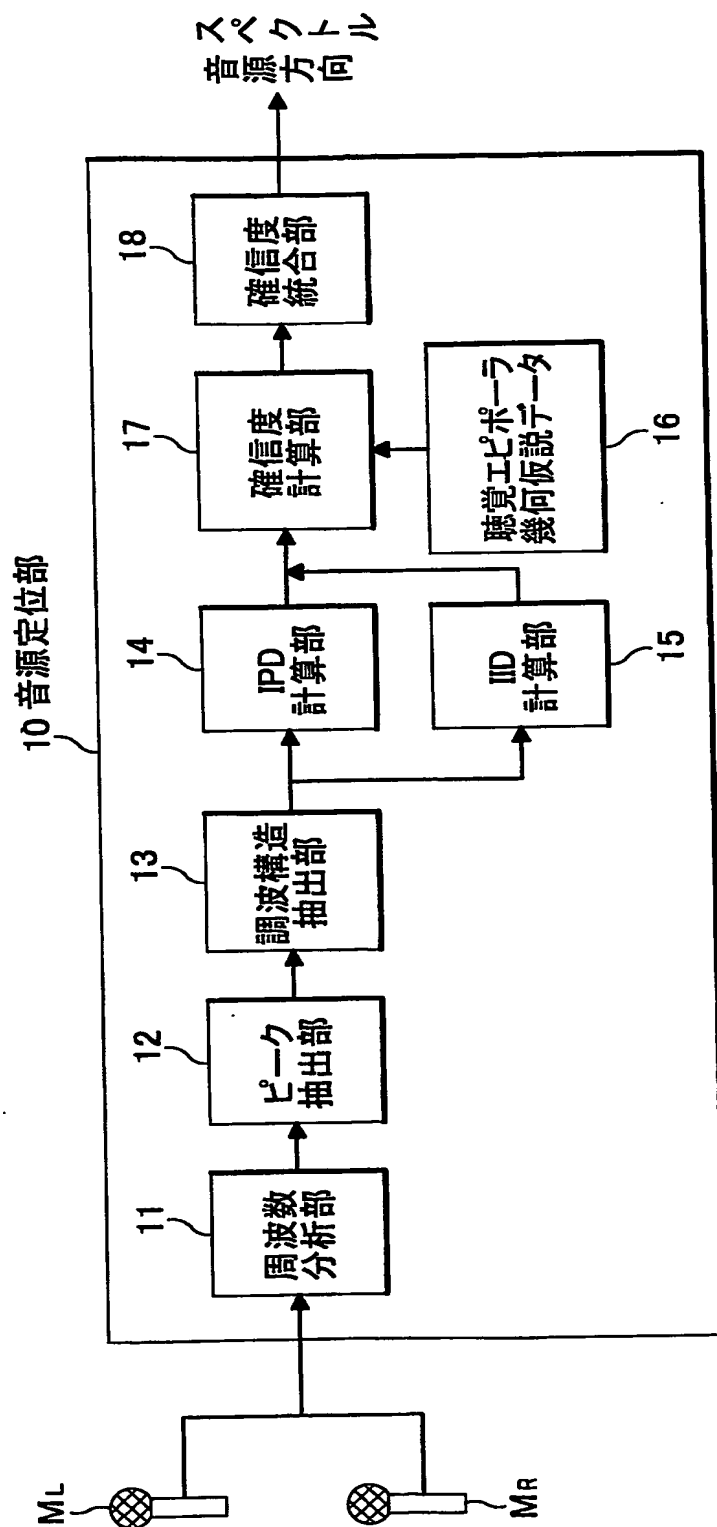


2 3 音響信号再合成部  
3 0 特徴抽出部  
4 0 音響モデル合成部  
4 9 音響モデル記憶部  
5 0 音声認識部  
M マイク

【書類名】 図面  
【図1】

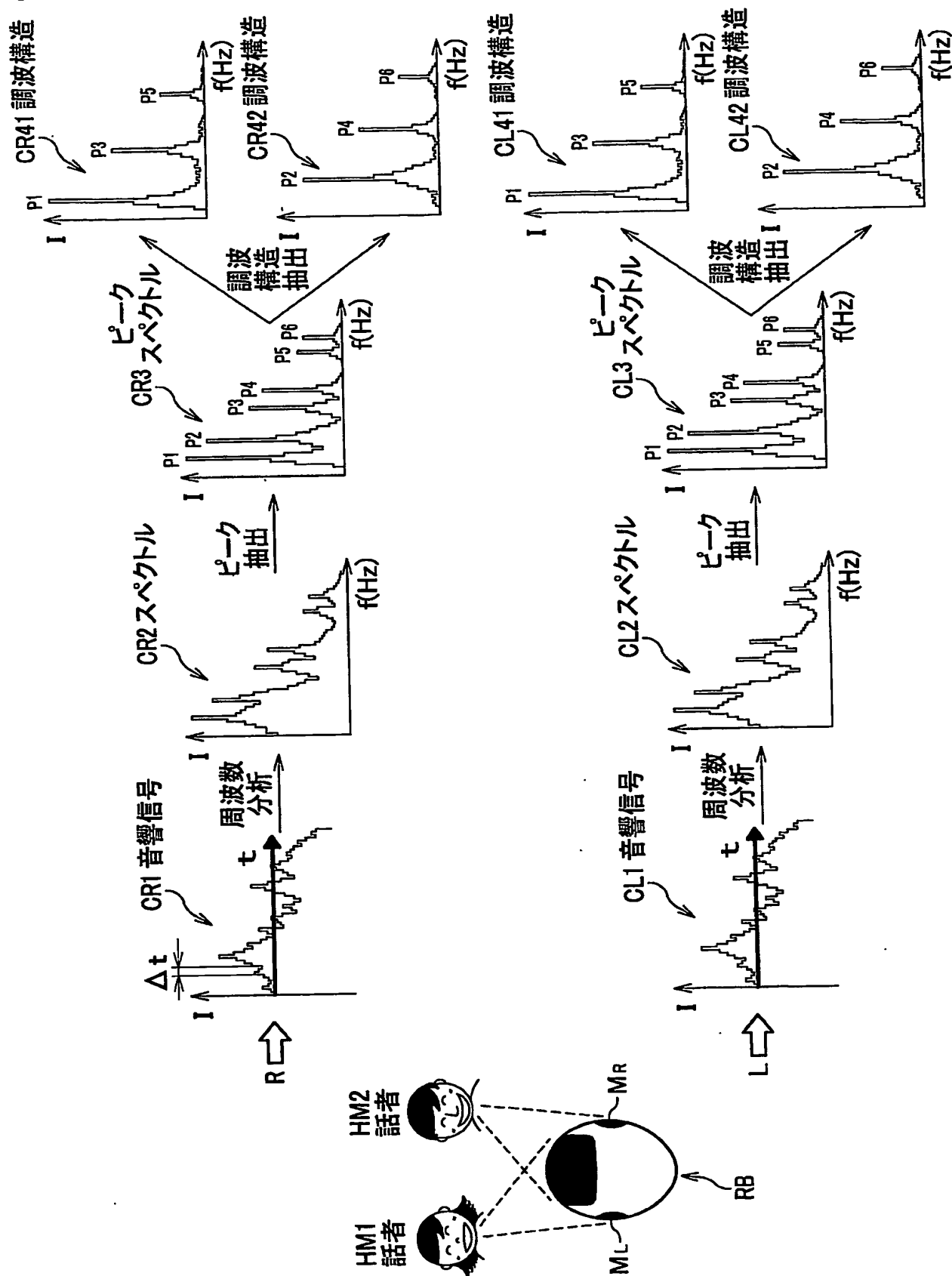


【図 2】

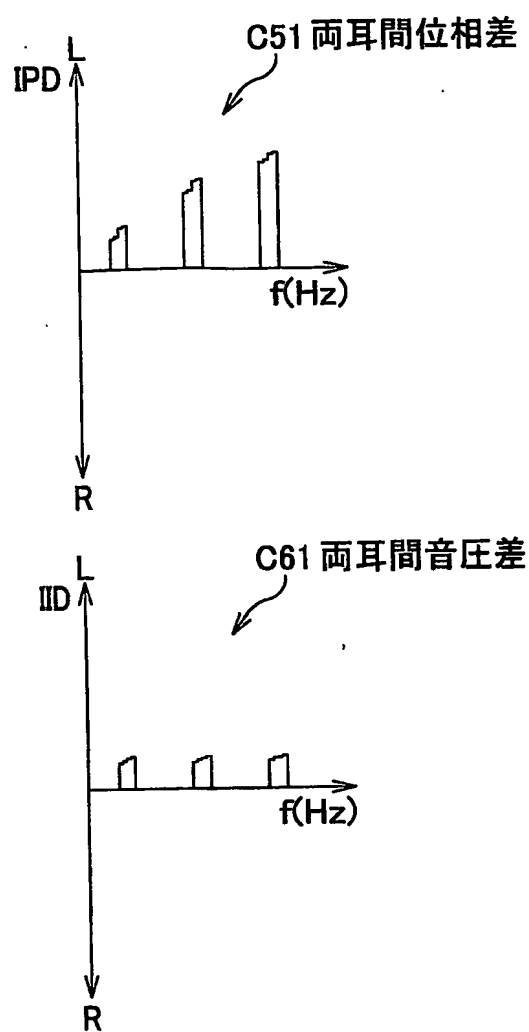




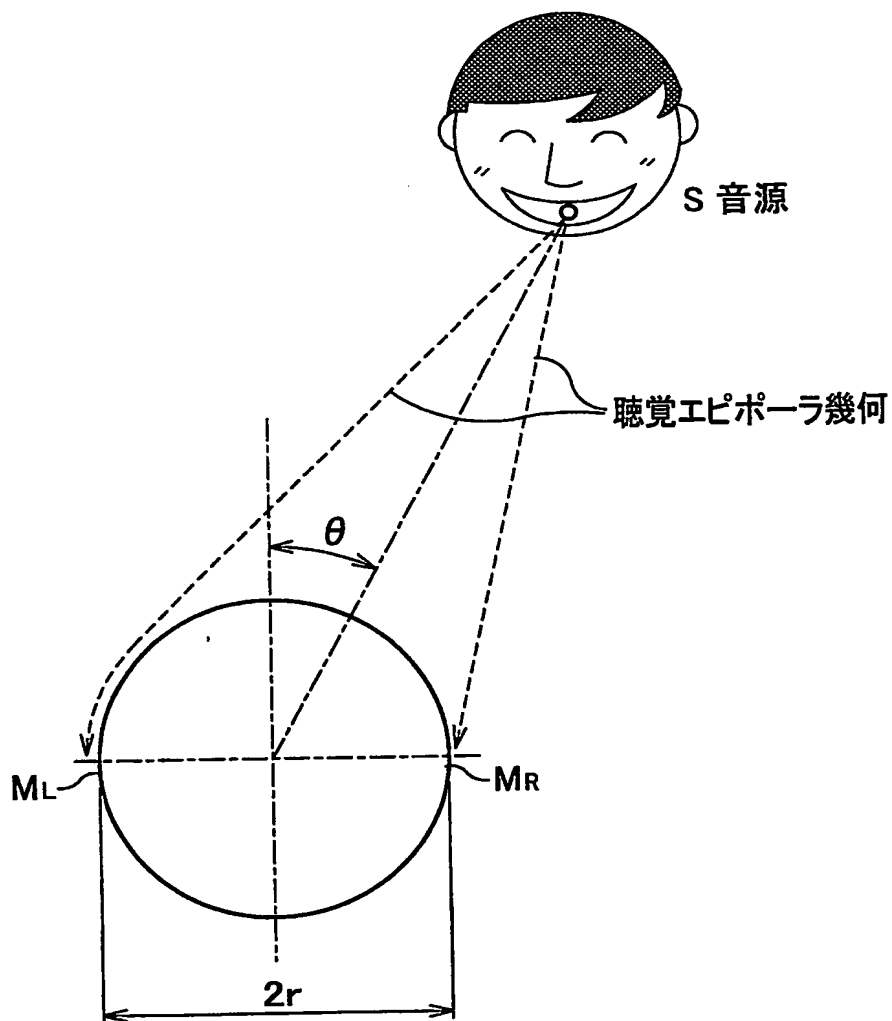
【図3】



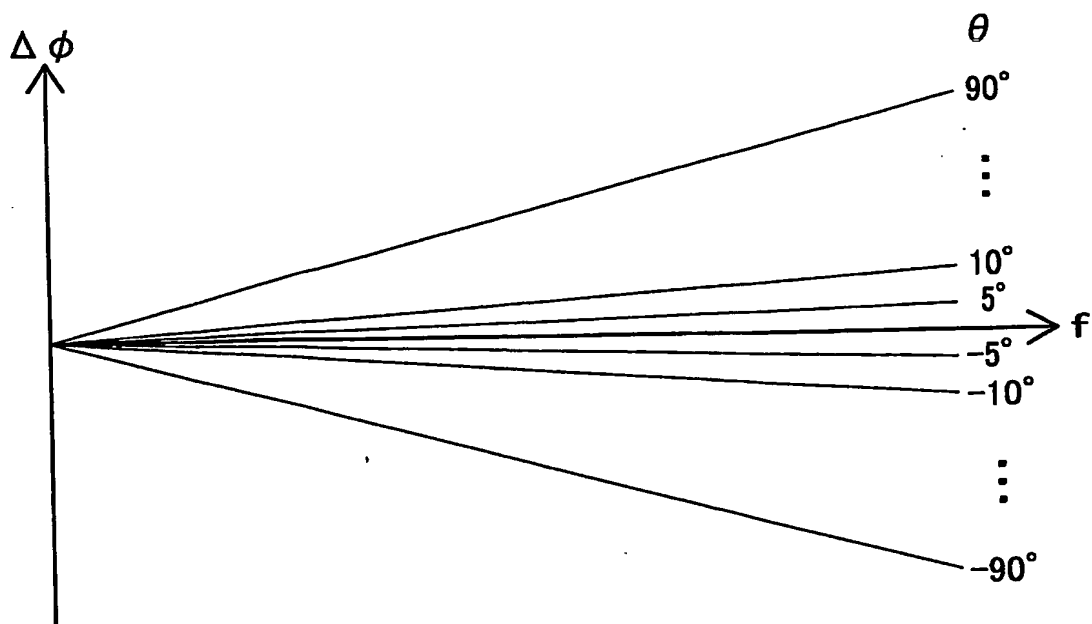
【図 4】



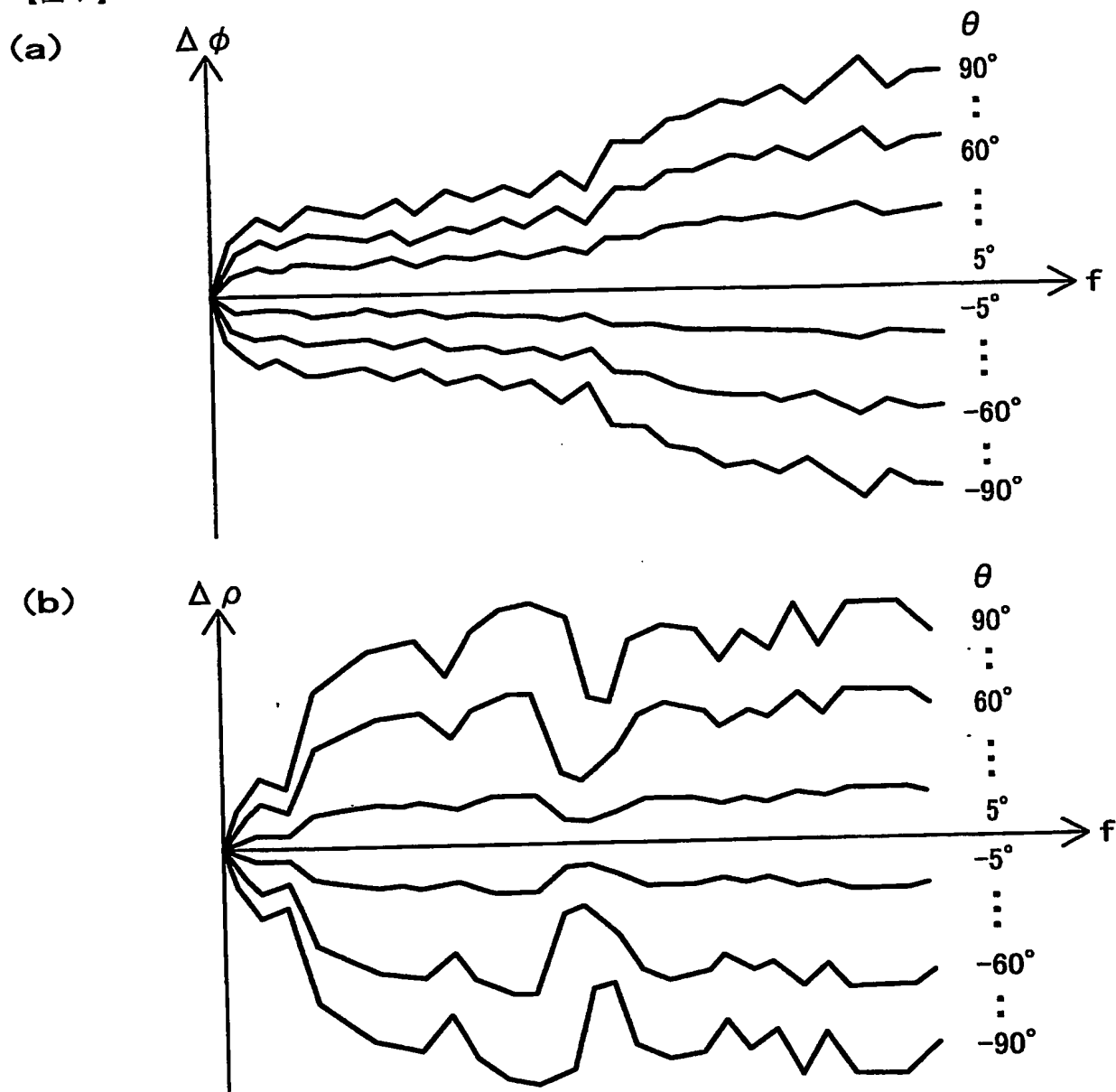
【図 5】



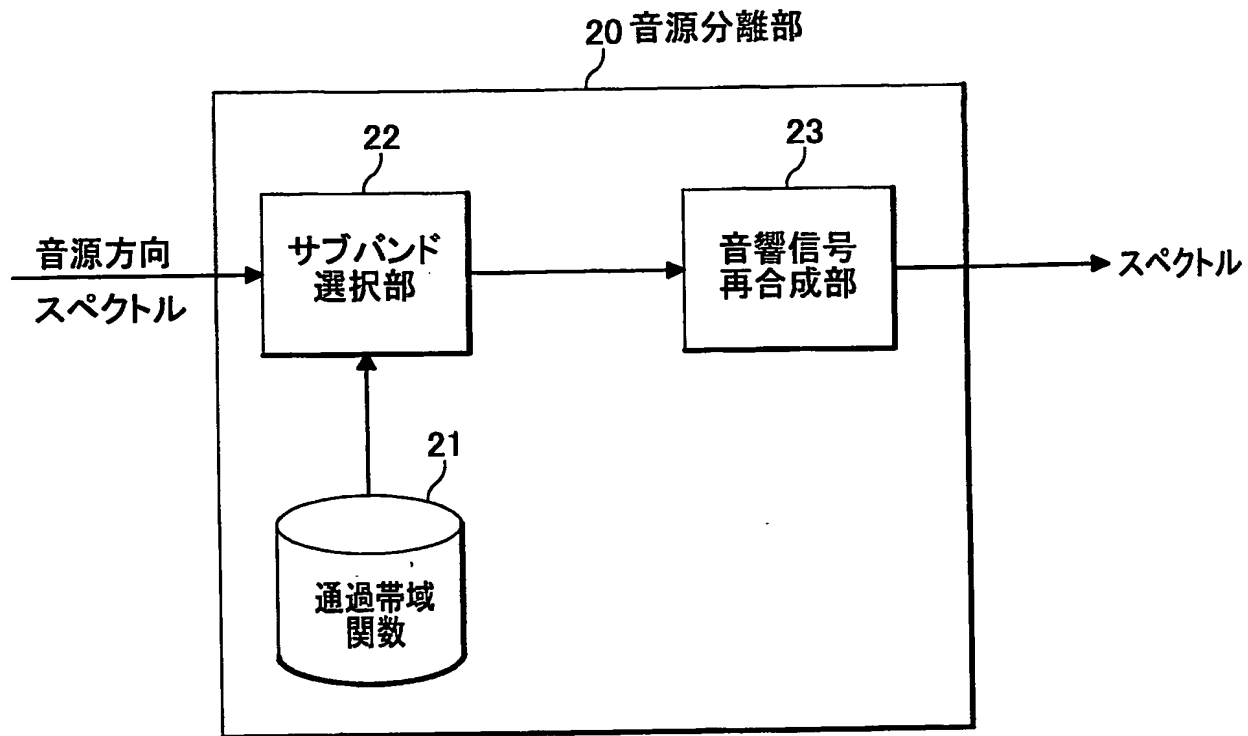
【図 6】



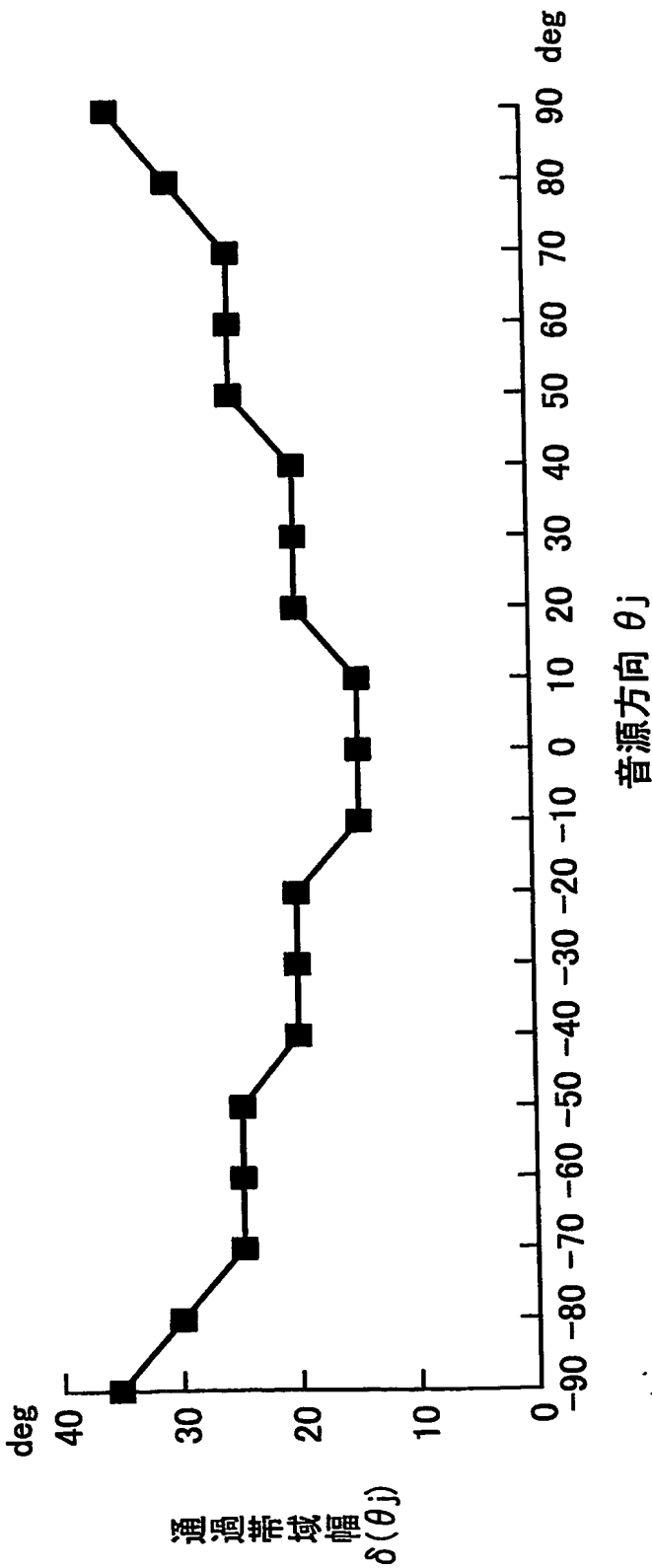
【図 7】



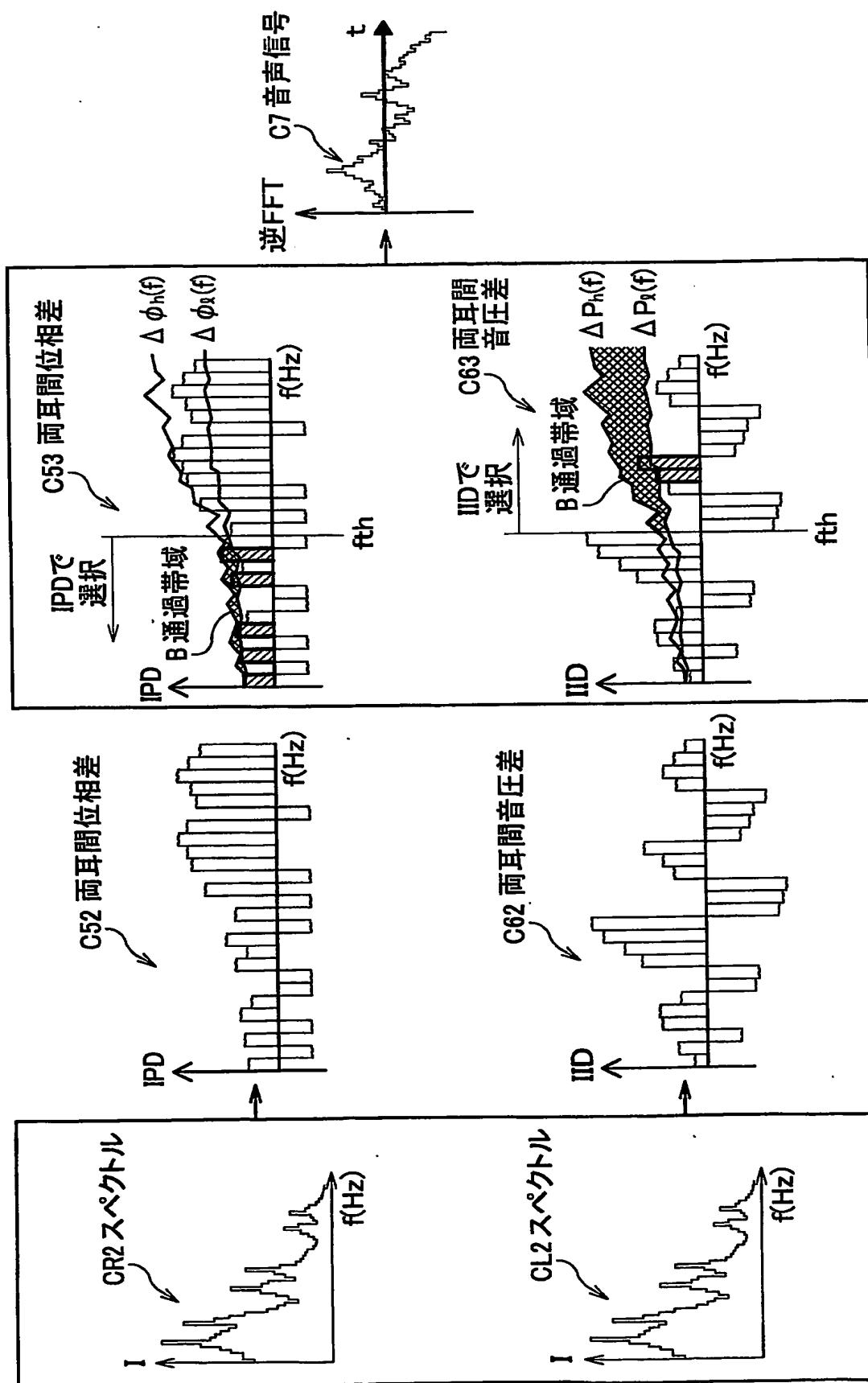
【図 8】



【図 9】

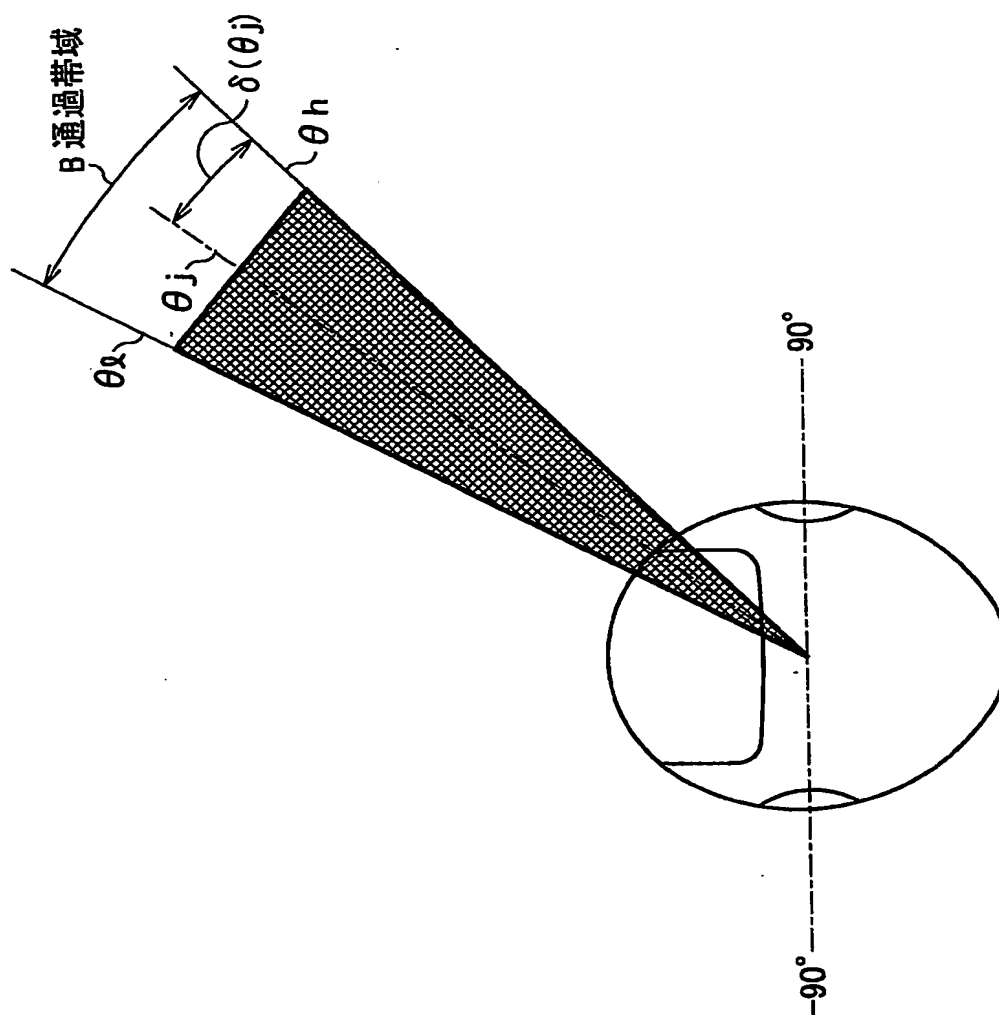


【図10】

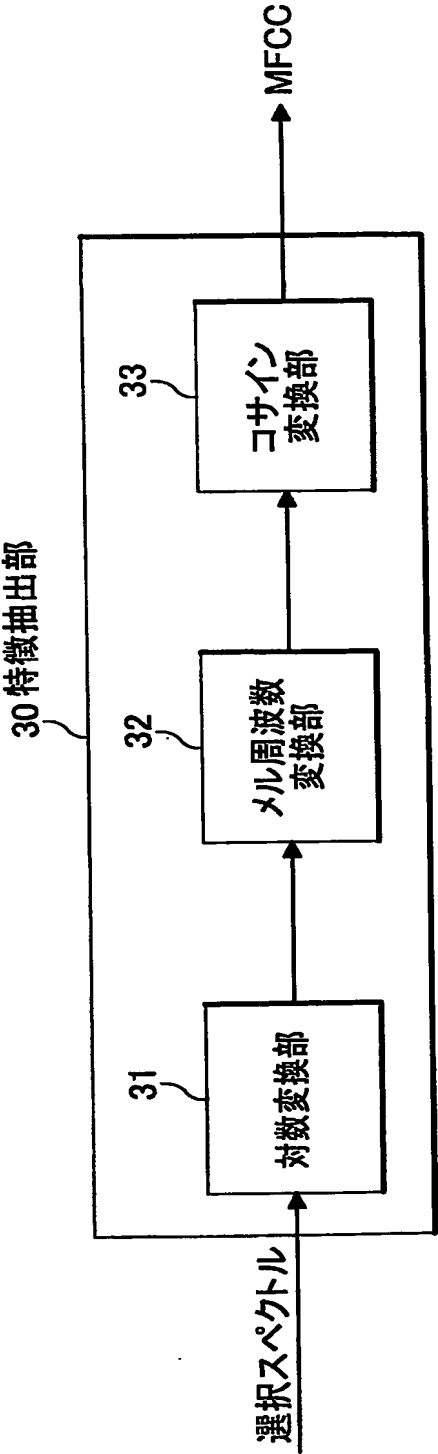




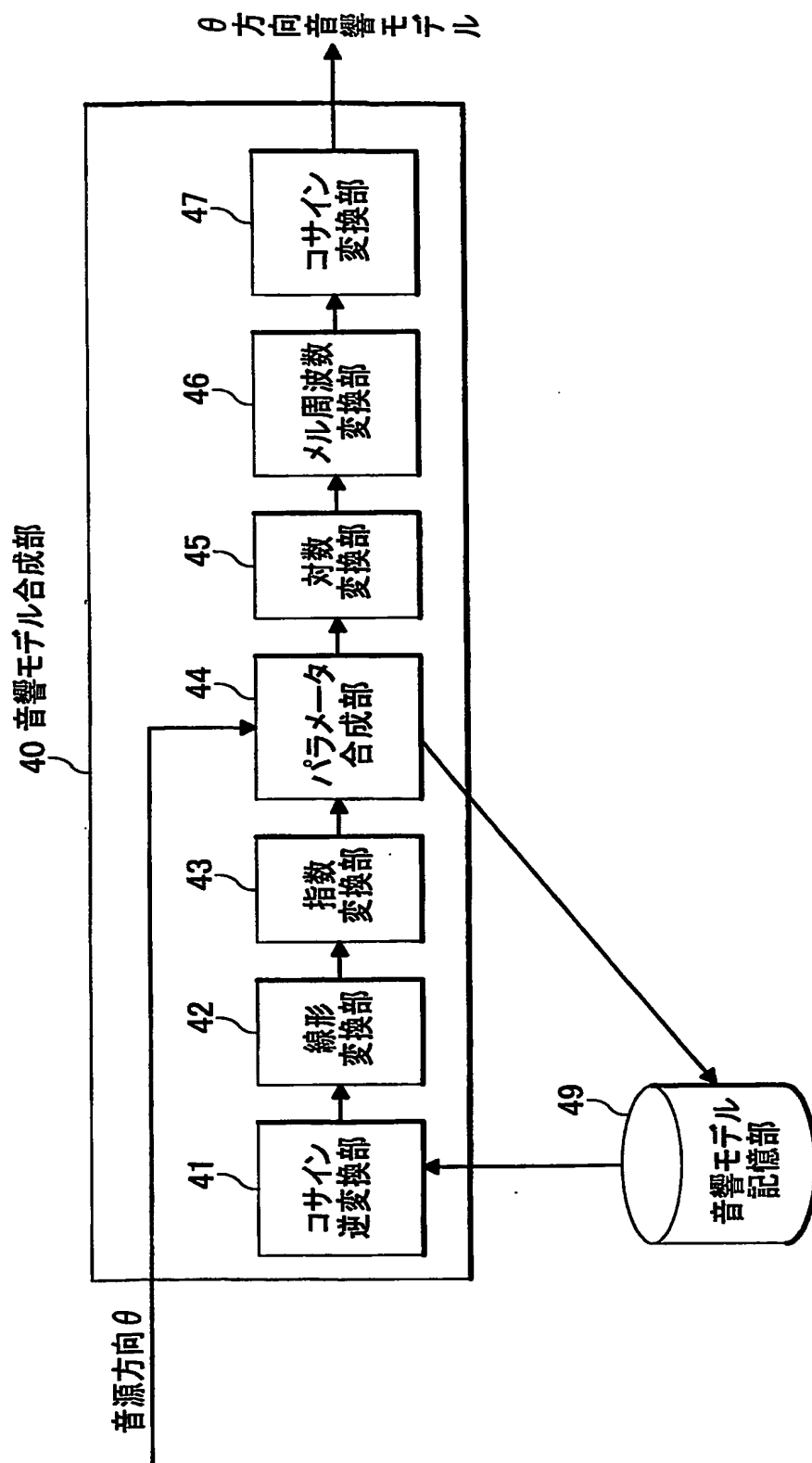
【図 11】



【図 12】



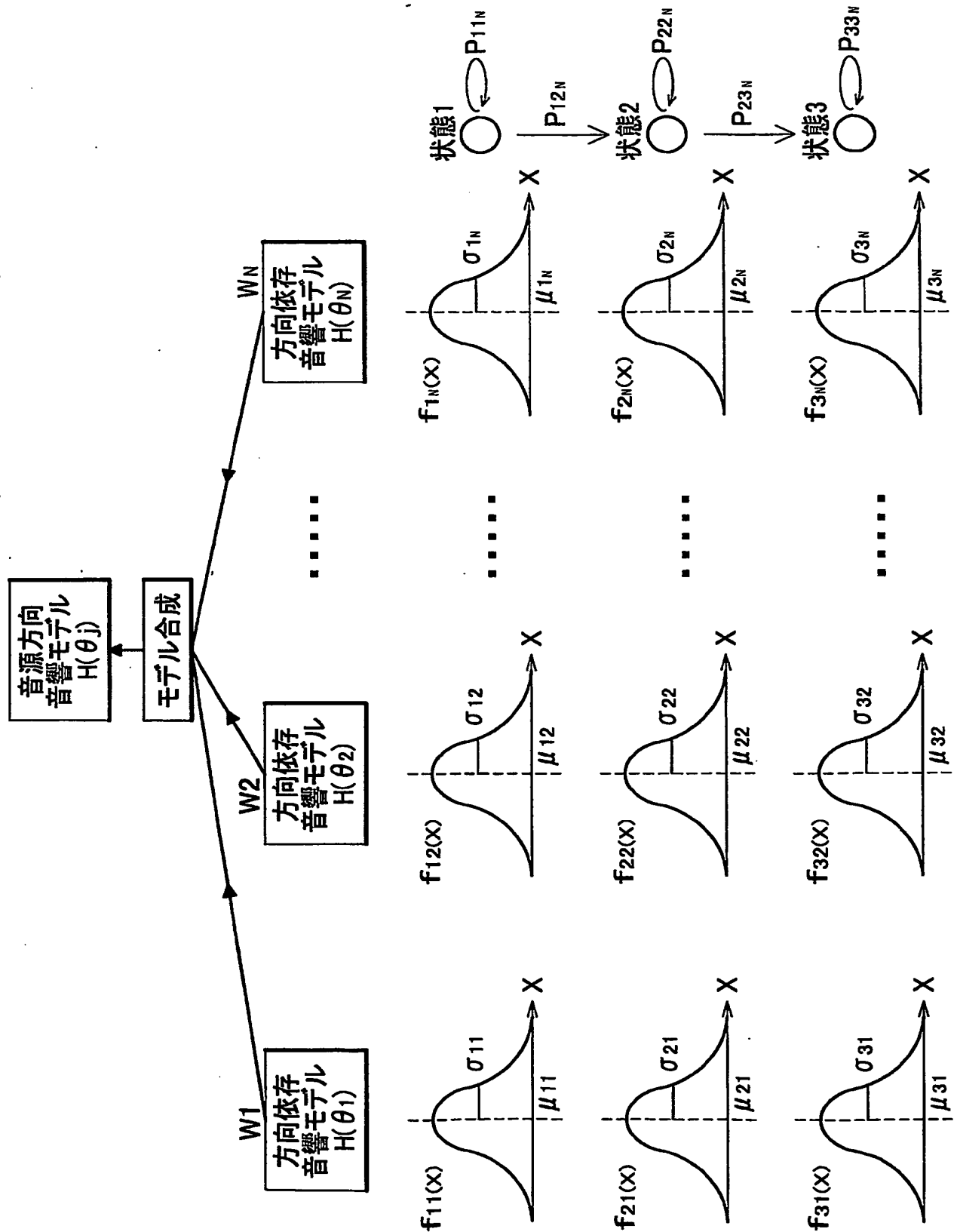
【図 13】



【図 14】

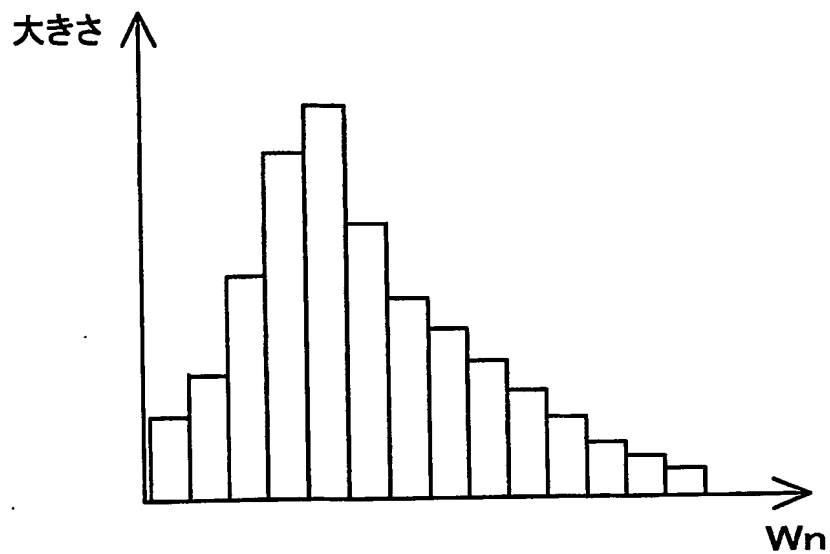
| 認識単位 | サブモデル              |
|------|--------------------|
| /a/  | $h(/a/, \theta_n)$ |
| /b/  | $h(/b/, \theta_n)$ |
| ⋮    | ⋮                  |
| m    | $h(m, \theta_n)$   |
| ⋮    | ⋮                  |

【図 15】



【図 16】

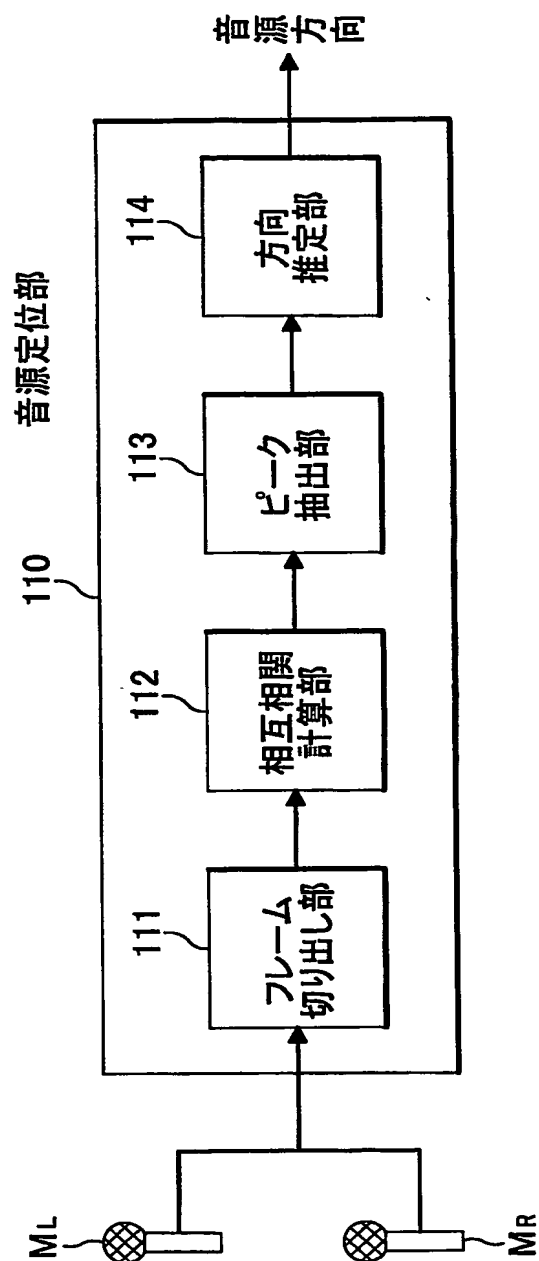
音源方向  $\theta_j$  の時の重み  $W_n$



【図 17】

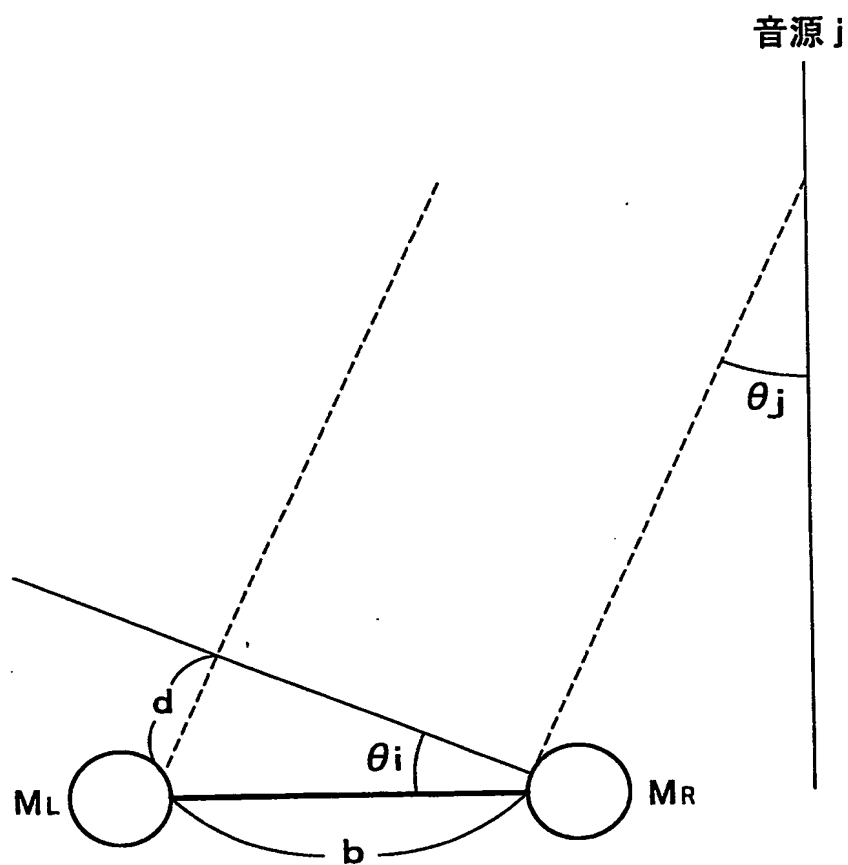
| 認識単位              | /a/      | /b/      | /c/      |
|-------------------|----------|----------|----------|
| $H(\theta_j)$     | /x/      | /y/      | /z/      |
| $H(\theta_{-90})$ | /x/      | /y/      | /c/      |
| $\vdots$          | $\vdots$ | $\vdots$ | $\vdots$ |
| $H(\theta_n)$     | /a/      | /y/      | /z/      |
| $\vdots$          | $\vdots$ | $\vdots$ | $\vdots$ |
| $H(\theta_{90})$  | /a/      | /y/      | /c/      |

【図 18】

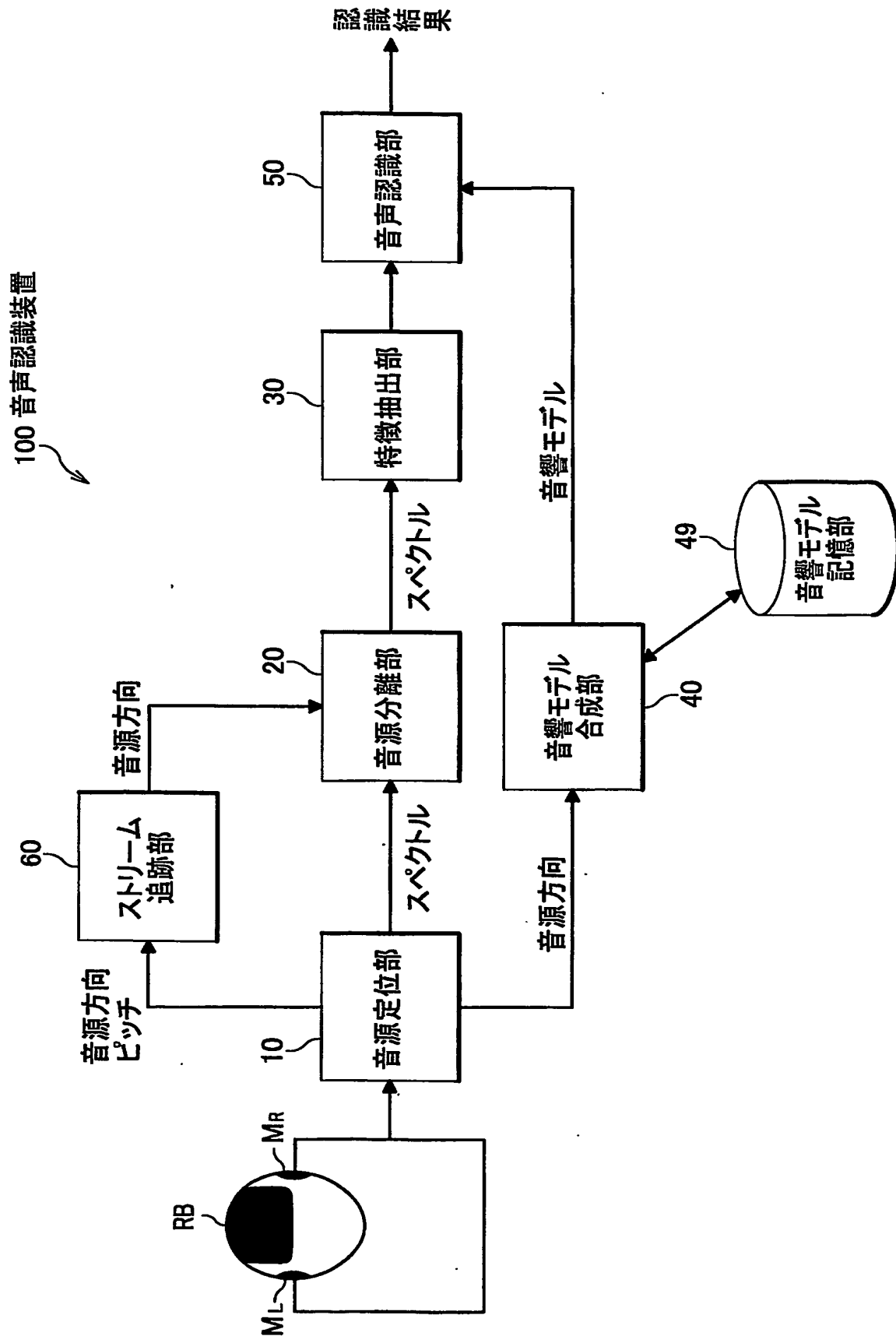




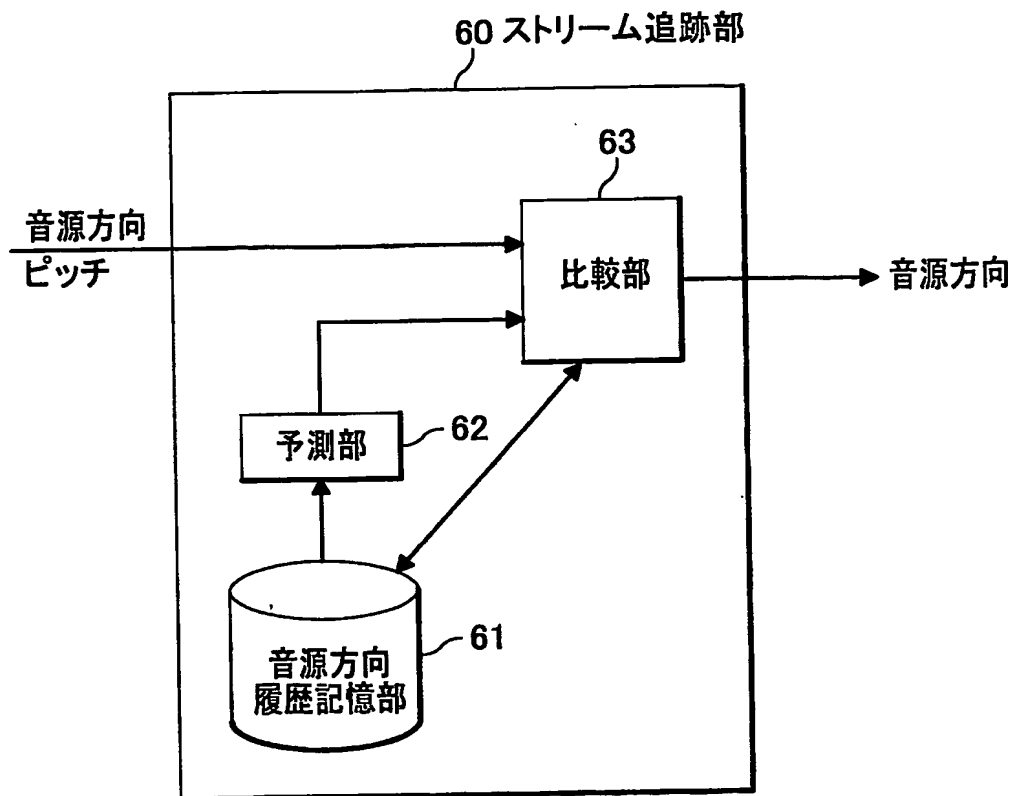
【図 19】



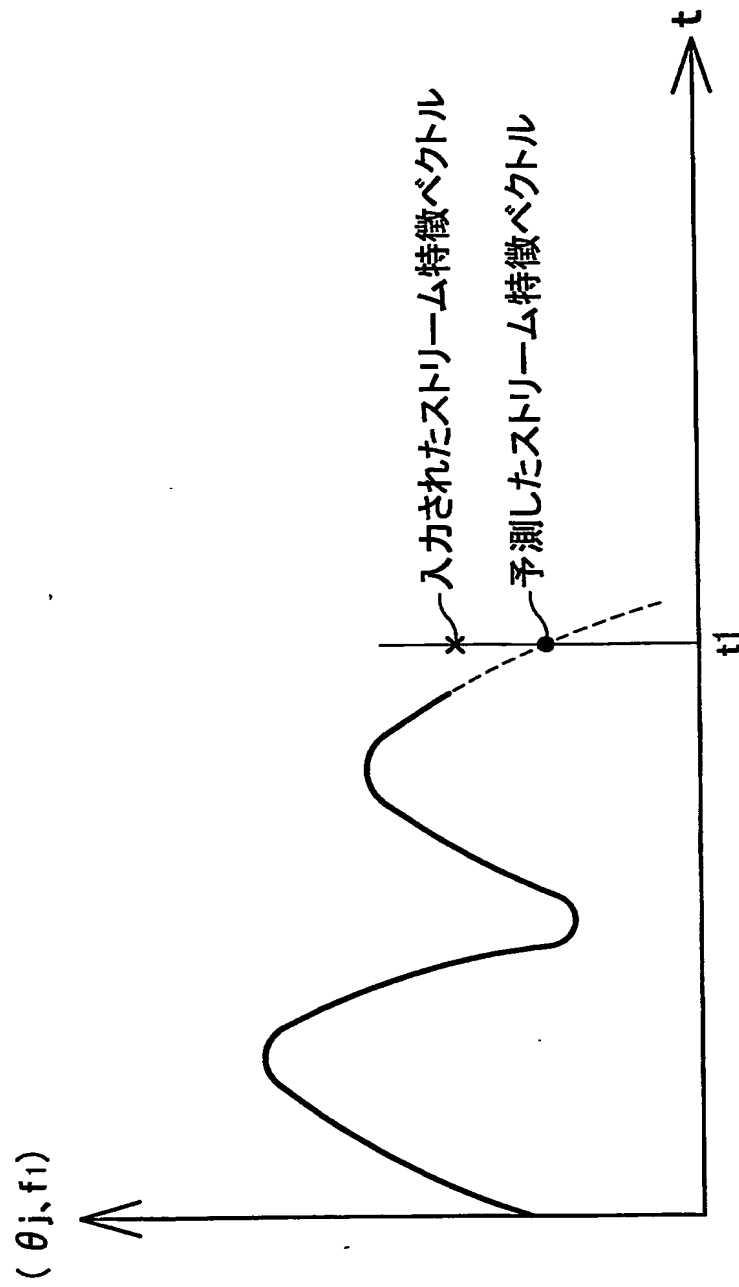
【図 20】



【図 21】



【図 22】



【書類名】 要約書

【要約】

【課題】 話者や、音声認識装置を搭載した移動体が移動しても高い精度で認識可能な音声認識装置を提供する。

【解決手段】 複数のマイクMが検出した音響信号から、特定の話者HM<sub>n</sub>の音声を認識して文字情報に変換する音声認識装置である。複数のマイクMが検出した音響信号に基づき、話者HM<sub>n</sub>の音源方向 $\theta_j$ を特定する音源定位部10と、音源方向 $\theta_j$ に基づき、話者HM<sub>n</sub>の音声信号を音響信号から分離する音源分離部20と、断続的な複数の方向に対応した方向依存音響モデル $H(\theta_n)$ を記憶した音響モデル記憶部49と、音源方向 $\theta_j$ の音響モデルを、音響モデル記憶部49の方向依存音響モデル $H(\theta_n)$ に基づいて求め、音響モデル記憶部49へ記憶させる音響モデル合成部40と、音響モデル合成部40が合成した音響モデルを使用して、音源分離部20が分離した音声信号の音声認識を行い、文字情報に変換する音声認識部50とを備える。

【選択図】 図1

特願 2 0 0 3 - 3 8 3 0 7 2

出 願 人 履 歴 情 報

識別番号

[ 0 0 0 0 0 5 3 2 6 ]

1. 変更年月日

1 9 9 0 年 9 月 6 日

[変更理由]

新規登録

住 所

東京都港区南青山二丁目 1 番 1 号

氏 名

本田技研工業株式会社

# Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP04/016883

International filing date: 12 November 2004 (12.11.2004)

Document type: Certified copy of priority document

Document details: Country/Office: JP  
Number: 2003-383072  
Filing date: 12 November 2003 (12.11.2003)

Date of receipt at the International Bureau: 20 January 2005 (20.01.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland  
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

**This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record.**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☒ **BLACK BORDERS**

☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**

☐ **FADED TEXT OR DRAWING**

☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**

☐ **SKEWED/SLANTED IMAGES**

☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**

☐ **GRAY SCALE DOCUMENTS**

☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**

☒ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**

☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**